

A dual consistent finite difference method with narrow stencil second derivative operators

Sofia Eriksson

Department of Mathematics, TU Darmstadt, Germany

Abstract

We study the numerical solutions of time-dependent systems of partial differential equations, focusing on the implementation of boundary conditions. The numerical method considered is a finite difference scheme constructed by high order summation by parts operators, combined with a boundary procedure using penalties (SBP-SAT).

Recently it was shown that SBP-SAT finite difference methods can yield superconvergent functional output if the boundary conditions are imposed such that the discretization is dual consistent. We generalize these results so that they include a broader range of boundary conditions and penalty parameters. The results are also generalized to hold for narrow-stencil second derivative operators. The derivations are supported by numerical experiments.

Keywords: Finite differences, summation by parts, simultaneous approximation term, dual consistency, superconvergence, functionals, narrow stencil

1 Introduction

In this paper we consider a summation by parts (SBP) finite difference method, which is combined with a penalty technique denoted simultaneous approximation term (SAT) for the boundary conditions. The main advantages of the SBP-SAT finite difference methods are high accuracy, computational efficiency and provable stability. For a background on the history and the newer developments of SBP-SAT, see [19, 6].

A discrete differential operator D_1 is said to be a SBP-operator if it can be factorized by the inverse of a positive definite matrix H and a difference operator Q , as specified later in equation (12). When H is diagonal, D_1 consists of a $2p$ -order accurate central difference approximation in the interior, but at the boundaries, the accuracy is limited to p th order. The global accuracy of the numerical solution can then be shown to be $p + 1$, see [19, 18].

In many applications functionals are of interest, sometimes they are even more important than the primary solution itself (one example is lift or drag coefficients in computational fluid dynamics). It could be expected that functionals computed from the numerical solution would have the same order of accuracy as the solution itself. However, recently Hicken and Zingg [9] showed that when computing the numerical solution in a dual consistent way, the order of accuracy of the output functional is higher than the FD solution itself, in fact, the full $2p$ accuracy can be recovered. Related papers are [10, 8] which includes interesting work on SBP operators as quadrature rules and error estimators for functional errors. Note that this kind of superconvergent behavior was already known for example for finite element and discontinuous Galerkin methods, but it had not been proven for finite difference schemes before, see [9]. Later Berg and Nordström [1, 2, 3] showed that the results hold also for time-dependent problems.

In [9, 8] and [1] boundary conditions of Dirichlet type are considered (in [9] Neumann boundary conditions are included but are rewritten on first order form), and in [2, 3] boundary conditions of far-field type are derived. In this paper, we generalize these results by deriving penalty parameters that yield dual consistency for all energy stable boundary conditions of Robin type (including the special cases Dirichlet and Neumann). In contrast to [2, 3], where the boundary conditions were adapted to get the penalty in a certain form, we adapt the penalty after the boundary conditions instead. Furthermore, we extend the results such that they hold also for narrow-stencil second derivative operators (sometimes also denoted compact second derivative operators), where the term narrow is used to define explicit finite difference schemes with a minimal stencil width. In fact, the results even carry over to narrow-stencil second derivatives operators for variable coefficients (of the type considered for example in [12]).

To keep things simple we consider linear problems in one spatial dimension, however, note that this is not due to a limitation of the method. In [9, 8] the extension to higher dimensions, curvilinear grids and non-linear problems are discussed and implemented for stationary problems and in [3] the theory is applied to the time-dependent Navier–Stokes and Euler equations in two dimensions.

The paper is organized as follows: In Section 2 we consider hyperbolic systems of partial differential equations and derive a family of SAT parameters which guarantees a stable and dual consistent discretization. Since higher order differential equations can always be rewritten as first order systems, this result directly leads to penalty parameters for parabolic problems, when using wide-stencil second derivative operators. Next, these parameters are generalized such that they hold also for narrow-stencil second derivative operators. This is all done in Section 3. In Section 4 a special aspect of the stability for the narrow operators is discussed. The derivations are then followed by examples and numerical simulations in Section 5 and a summary is given in Section 6.

1.1 Preliminaries

We consider time-dependent partial differential equations (PDE) as

$$\mathcal{U}_t + \mathcal{L}(\mathcal{U}) = \mathcal{F}, \quad t \in [0, T], \quad x \in \Omega, \quad (1)$$

where \mathcal{L} represents a linear, spatial differential operator and $\mathcal{F}(x, t)$ is a forcing function. For simplicity, we will assume that the sought solution $\mathcal{U}(x, t)$ satisfies homogeneous initial and boundary conditions. To derive the dual equations we follow [9, 1, 2] and pose the problem in a variational framework: Given a functional $\mathcal{J}(\mathcal{U}) = \langle \mathcal{G}, \mathcal{U} \rangle$, where $\mathcal{G}(x, t)$ is smooth weight function and where $\langle \mathcal{G}, \mathcal{U} \rangle = \int_{\Omega} \mathcal{G}^T \mathcal{U} dx$ refers to the standard L^2 inner product, we seek a function $\mathcal{V}(x, t)$ such that $\mathcal{J}(\mathcal{U}) = \mathcal{J}^*(\mathcal{V}) = \langle \mathcal{V}, \mathcal{F} \rangle$. This defines the dual problem as

$$\mathcal{V}_{\tau} + \mathcal{L}^*(\mathcal{V}) = \mathcal{G}, \quad \tau \in [0, T], \quad x \in \Omega, \quad (2)$$

where \mathcal{L}^* is the adjoint operator, given by $\langle \mathcal{V}, \mathcal{L}\mathcal{U} \rangle = \langle \mathcal{L}^*\mathcal{V}, \mathcal{U} \rangle$, and where \mathcal{V} also satisfies homogeneous initial and boundary conditions. Note that the dual problem actually goes "backward" in time; the expression in (2) is obtained using the transformation $\tau = T - t$.

Let U and V be discrete vectors approximating \mathcal{U} and \mathcal{V} , respectively, and let F and G be projections of \mathcal{F} and \mathcal{G} onto a spatial grid. We discretize (1) using a stable and consistent SBP-SAT scheme, leading to

$$U_t + LU = F, \quad t \in [0, T]. \quad (3)$$

The SBP scheme has an associated matrix H which defines a discrete inner product, as $\langle G, U \rangle_H = G^T H U$ (when \mathcal{U} is vector-valued, H must be replaced by \bar{H} , which is defined later in the paper). Now the discrete adjoint operator is given by $L^* = H^{-1} L^T H$, since this leads to $\langle V, LU \rangle_H = \langle L^* V, U \rangle_H$ which mimics the continuous relation above.

If L^* happens to be a consistent approximation of \mathcal{L}^* , then the discretization (3) is said to be *dual consistent* (if considering the stationary case) or *spatially dual consistent*, see [9, 1] respectively. When (3) is a stable and dual consistent discretization of (1), then the linear functional $J(U) = \langle G, U \rangle_H$ is a $2p$ -order accurate approximation of $\mathcal{J}(\mathcal{U})$, that is $J(U) = \mathcal{J}(\mathcal{U}) + \mathcal{O}(h^{2p})$, and we thus have superconvergent functional output. To obtain such high accuracy it is necessary with compatible and sufficiently smooth data, see [9] for more details.

2 Hyperbolic systems

We start by considering a hyperbolic system of PDEs of reaction-advection type, namely

$$\begin{aligned} \mathcal{U}_t + \mathcal{R}\mathcal{U} + \mathcal{A}\mathcal{U}_x &= \mathcal{F}, & x \in [x_L, x_R], \\ \mathcal{B}_L\mathcal{U} &= g_L, & x = x_L, \\ \mathcal{B}_R\mathcal{U} &= g_R, & x = x_R, \end{aligned} \tag{4}$$

valid for $t \geq 0$ and augmented with initial data $\mathcal{U}(x, 0) = \mathcal{U}_0(x)$. We let \mathcal{R} and \mathcal{A} be real-valued, symmetric $n \times n$ matrices. Further, \mathcal{R} is positive semi-definite, that is $\mathcal{R} \geq 0$. The operators \mathcal{B}_L and \mathcal{B}_R define the form of the boundary conditions and their properties are specified in (10) below. The forcing function $\mathcal{F}(x, t)$, the initial data $\mathcal{U}_0(x)$ and the boundary data $g_L(t)$ and $g_R(t)$ are assumed to be compatible and sufficiently smooth such that the solution $\mathcal{U}(x, t)$ exists. We will refer to (4) as our primal problem.

2.1 Well-posedness using the energy method

We call (4) well-posed if it has a unique solution and is stable. Existence is guaranteed by using the right number of boundary conditions, and uniqueness then follows from the stability, [15, 7]. Next we show stability, using the energy method.

The PDE in the first row of (4) is multiplied by \mathcal{U}^T from the left and integrated over the domain $\Omega = [x_L, x_R]$. Using integration by parts we obtain

$$\frac{d}{dt} \|\mathcal{U}\|^2 + 2\langle \mathcal{U}, \mathcal{R}\mathcal{U} \rangle = 2\langle \mathcal{U}, \mathcal{F} \rangle + \text{BT}_L + \text{BT}_R \tag{5}$$

where $\|\mathcal{U}\|^2 = \langle \mathcal{U}, \mathcal{U} \rangle = \int_{x_L}^{x_R} \mathcal{U}^T \mathcal{U} dx$ and where

$$\text{BT}_L = \mathcal{U}^T \mathcal{A} \mathcal{U} \big|_{x_L}, \quad \text{BT}_R = -\mathcal{U}^T \mathcal{A} \mathcal{U} \big|_{x_R}.$$

To bound the growth of the solution, we must ensure that the boundary conditions make BT_L and BT_R non-positive for zero data. We consider the matrix \mathcal{A} above and assume that we have found a factorization such that

$$\mathcal{A} = Z\Delta Z^T, \quad \Delta = \begin{bmatrix} \Delta_+ & & \\ & \Delta_0 & \\ & & \Delta_- \end{bmatrix}, \quad Z = [Z_+, Z_0, Z_-], \tag{6}$$

where Z is non-singular. The parts of Δ are arranged such that $\Delta_+ > 0$, $\Delta_0 = 0$ and $\Delta_- < 0$. According to Sylvester's law of inertia, the matrices \mathcal{A} and Δ have the same number of positive (n_+), negative (n_-) and zero (n_0) eigenvalues (for a non-singular

Z), where $n = n_+ + n_0 + n_-$. To bound the terms BT_L and BT_R , we have to give n_+ boundary conditions at $x = x_L$ and n_- boundary conditions at $x = x_R$. We note that

$$\mathcal{A} = Z_+ \Delta_+ Z_+^T + Z_- \Delta_- Z_-^T, \quad (7)$$

which gives

$$\text{BT}_L = \mathcal{U}^T (Z_+ \Delta_+ Z_+^T + Z_- \Delta_- Z_-^T) \mathcal{U}|_{x_L}, \quad \text{BT}_R = -\mathcal{U}^T (Z_+ \Delta_+ Z_+^T + Z_- \Delta_- Z_-^T) \mathcal{U}|_{x_R}$$

where $Z_+^T \mathcal{U}$ represents the right-going variables (ingoing at the left boundary), and $Z_-^T \mathcal{U}$ represents the left-going variables (ingoing at the right boundary). The ingoing variables are given data in terms of known functions and outgoing variables, as

$$Z_+^T \mathcal{U}|_{x_L} = \tilde{g}_L - R_L Z_-^T \mathcal{U}|_{x_L}, \quad Z_-^T \mathcal{U}|_{x_R} = \tilde{g}_R - R_R Z_+^T \mathcal{U}|_{x_R}, \quad (8)$$

where \tilde{g}_L, \tilde{g}_R are the known data and where the matrices R_L and R_R must be sufficiently small. Using the boundary conditions in (8), the boundary terms BT_L and BT_R become

$$\begin{aligned} \text{BT}_L &= \mathcal{U}^T Z_- (\Delta_- + R_L^T \Delta_+ R_L) Z_-^T \mathcal{U}|_{x_L} - 2\tilde{g}_L^T \Delta_+ R_L Z_-^T \mathcal{U}|_{x_L} + \tilde{g}_L^T \Delta_+ \tilde{g}_L \\ \text{BT}_R &= -\mathcal{U}^T Z_+ (\Delta_+ + R_R^T \Delta_- R_R) Z_+^T \mathcal{U}|_{x_R} + 2\tilde{g}_R^T \Delta_- R_R Z_+^T \mathcal{U}|_{x_R} - \tilde{g}_R^T \Delta_- \tilde{g}_R. \end{aligned} \quad (9)$$

We define

$$\mathcal{C}_L = \Delta_- + R_L^T \Delta_+ R_L, \quad \mathcal{C}_R = -\Delta_+ - R_R^T \Delta_- R_R$$

and note that if $\mathcal{C}_L, \mathcal{C}_R \leq 0$, the boundary terms in (9) will be non-positive for zero data. By integrating (5) in time we can now obtain a bound on $\|\mathcal{U}\|^2$. With boundary conditions on the form (8), we also know that the correct number of boundary conditions are specified at each boundary, which yields existence. Our problem is thus well-posed.

To relate the original boundary conditions in (4) to the ones in (8), we let

$$\mathcal{B}_L = P_L (Z_+^T + R_L Z_-^T), \quad \mathcal{B}_R = P_R (Z_-^T + R_R Z_+^T), \quad (10)$$

where P_L and P_R are invertible scaling and/or permutation matrices. The data in (8) is identified as $\tilde{g}_L = P_L^{-1} g_L$ and $\tilde{g}_R = P_R^{-1} g_R$. We assume that the boundary conditions in (4) are properly chosen such that R_L and R_R are sufficiently small and hence $\mathcal{C}_L, \mathcal{C}_R \leq 0$.

Remark 2.1. Note that the energy method is a sufficient but not necessary condition for stability and that it is rather restrictive with respect to the admissible boundary conditions. By rescaling the problem we could allow R_L and R_R to be larger, see [11, 7]. We will not consider this complication but simply require that $\mathcal{C}_L \leq 0, \mathcal{C}_R \leq 0$.

Remark 2.2. In the homogeneous case, with boundary conditions such that $\mathcal{C}_L, \mathcal{C}_R \leq 0$, the growth rate in (5) becomes $\frac{d}{dt} \|\mathcal{U}\|^2 \leq 0$. Integrating this in time we obtain the energy estimate $\|\mathcal{U}\|^2 \leq \|\mathcal{U}_0\|^2$ and (4) is well-posed. Since (4) is an one-dimensional hyperbolic problem it is also possible to show strong well-posedness, i.e. that $\|\mathcal{U}\|$ is bounded by the data g_L, g_R, \mathcal{F} and \mathcal{U}_0 . See [11, 7] for different definitions of well-posedness.

2.2 The semi-discrete problem

We discretize in space using $N + 1$ equidistant grid points $x_i = x_L + hi$, where $h = (x_R - x_L)/N$ and $i = 0, 1, \dots, N$. The semi-discrete scheme approximating (4) is written

$$U_t + (I_N \otimes \mathcal{R})U + (D_1 \otimes \mathcal{A})U = F + (H^{-1}e_0 \otimes \Sigma_0)(\mathcal{B}_L U_0 - g_L) \\ + (H^{-1}e_N \otimes \Sigma_N)(\mathcal{B}_R U_N - g_R), \quad (11)$$

where $U = [U_0^T, U_1^T, \dots, U_N^T]^T$ is a vector of length $n(N + 1)$, such that $U_i(t) \approx \mathcal{U}(x_i, t)$, and where $F_i(t) = \mathcal{F}(x_i, t)$. The symbol \otimes refers to the Kronecker product. The finite difference operator D_1 approximates $\partial/\partial x$ and satisfies the SBP-properties

$$D_1 = H^{-1}Q, \quad H = H^T > 0, \quad Q + Q^T = E_N - E_0 \quad (12)$$

where $E_0 = e_0 e_0^T$, $E_N = e_N e_N^T$ and $e_0 = [1, 0, \dots, 0]^T$ and $e_N = [0, \dots, 0, 1]^T$. Note that $U_0 = (e_0^T \otimes I_n)U$ and $U_N = (e_N^T \otimes I_n)U$. By I_N and I_n we refer to identity matrices of size $N+1$ and n , respectively. The boundary conditions are imposed using the SAT technique which is a penalty method. The penalty parameters Σ_0 and Σ_N in (11) are at this point unknown, but are derived in the next subsections and presented in Theorem 2.6.

In this paper, we require that H is diagonal, and in this case D_1 consists of a $2p$ -order accurate central difference approximation in the interior and one-sided, p -order accurate approximations at the boundaries. Examples of SBP operators can be found in [17, 13]. For more details about SBP-SAT, see [18] and references therein.

2.3 Numerical stability using the energy method

Just as in the continuous case we use the energy method to show stability. We multiply (11) by $U^T \bar{H}$ from the left, where $\bar{H} = H \otimes I_n$, and then add the transpose of the result. Thereafter using the SBP-properties in (12) we obtain

$$\frac{d}{dt} \|U\|_H^2 + 2U^T (H \otimes \mathcal{R})U = 2\langle U, F \rangle_H + \text{BT}_L^{Disc.} + \text{BT}_R^{Disc.},$$

where $\|U\|_H^2 = \langle U, U \rangle_H = U^T \bar{H}U$ is the discrete L^2 -norm and where

$$\text{BT}_L^{Disc.} = U_0^T (\mathcal{A} + \Sigma_0 \mathcal{B}_L + \mathcal{B}_L^T \Sigma_0^T) U_0 - U_0^T \Sigma_0 g_L - g_L^T \Sigma_0^T U_0, \\ \text{BT}_R^{Disc.} = U_N^T (-\mathcal{A} + \Sigma_N \mathcal{B}_R + \mathcal{B}_R^T \Sigma_N^T) U_N - U_N^T \Sigma_N g_R - g_R^T \Sigma_N^T U_N. \quad (13)$$

We define $C_0 = \mathcal{A} + \Sigma_0 \mathcal{B}_L + \mathcal{B}_L^T \Sigma_0^T$ and $C_N = -\mathcal{A} + \Sigma_N \mathcal{B}_R + \mathcal{B}_R^T \Sigma_N^T$. For stability $\text{BT}_L^{Disc.}$ and $\text{BT}_R^{Disc.}$ must be non-positive for zero boundary data, i.e. $C_0 \leq 0$ and $C_N \leq 0$. We make the following ansätze for the penalty parameters:

$$\Sigma_0 = (Z_+ \Pi_0 + Z_- \Gamma_0) P_L^{-1}, \quad \Sigma_N = (Z_+ \Gamma_N + Z_- \Pi_N) P_R^{-1}. \quad (14)$$

Taking the left boundary as example and using (7), (10) and (14) we obtain

$$C_0 = \begin{bmatrix} Z_+^T \\ Z_-^T \end{bmatrix}^T \begin{bmatrix} \Delta_+ + \Pi_0 + \Pi_0^T & \Pi_0 R_L + \Gamma_0^T \\ \Gamma_0 + R_L^T \Pi_0^T & \Delta_- + \Gamma_0 R_L + R_L^T \Gamma_0^T \end{bmatrix} \begin{bmatrix} Z_+^T \\ Z_-^T \end{bmatrix}. \quad (15)$$

2.4 The dual problem

Given the functional $\mathcal{J}(\mathcal{U}) = \langle \mathcal{G}, \mathcal{U} \rangle$, the dual problem of (4) is

$$\begin{aligned} \mathcal{V}_\tau + \mathcal{R}\mathcal{V} - \mathcal{A}\mathcal{V}_x &= \mathcal{G}, & x \in [x_L, x_R], \\ \widetilde{\mathcal{B}}_L \mathcal{V} &= \widetilde{g}_L, & x = x_L, \\ \widetilde{\mathcal{B}}_R \mathcal{V} &= \widetilde{g}_R, & x = x_R, \end{aligned} \quad (16)$$

which holds for $\tau \geq 0$ and is complemented with the initial condition $\mathcal{V}(x, 0) = \mathcal{V}_0(x)$. The boundary operators in (16) have the form

$$\widetilde{\mathcal{B}}_L = \widetilde{P}_L(Z_-^T + \widetilde{R}_L Z_+^T), \quad \widetilde{\mathcal{B}}_R = \widetilde{P}_R(Z_+^T + \widetilde{R}_R Z_-^T), \quad (17)$$

where \widetilde{P}_L and \widetilde{P}_R are arbitrary invertible matrices and \widetilde{R}_L and \widetilde{R}_R depend on the primal boundary conditions as

$$\widetilde{R}_L = -\Delta_-^{-1} R_L^T \Delta_+, \quad \widetilde{R}_R = -\Delta_+^{-1} R_R^T \Delta_-. \quad (18)$$

The claim that (16), (17) and (18) describes the dual problem is motivated below: Using the notation in (1) and (2) we identify the spatial operators of (4) and (16) as

$$\mathcal{L} = \mathcal{R} + \mathcal{A} \frac{\partial}{\partial x}, \quad \mathcal{L}^* = \mathcal{R} - \mathcal{A} \frac{\partial}{\partial x}, \quad (19)$$

respectively. For (16) to be the dual problem of (4), \mathcal{L} and \mathcal{L}^* must fulfill the relation $\langle \mathcal{V}, \mathcal{L}\mathcal{U} \rangle = \langle \mathcal{L}^*\mathcal{V}, \mathcal{U} \rangle$. Using integration by parts we obtain

$$\langle \mathcal{V}, \mathcal{L}\mathcal{U} \rangle = \langle \mathcal{L}^*\mathcal{V}, \mathcal{U} \rangle + [\mathcal{V}^T \mathcal{A}\mathcal{U}]_{x_L}^{x_R}$$

and we see that $\mathcal{V}^T \mathcal{A}\mathcal{U}$ must be zero at both boundaries. (The boundary conditions for the dual problem are defined as the minimal set of homogeneous conditions such that all boundary terms vanish after that the homogeneous boundary conditions for the primal problem have been applied, see [1].) Using the boundary conditions of the primal problem, (8), followed by the dual boundary conditions, (16), (17), yields (for zero data)

$$\begin{aligned} \mathcal{V}^T \mathcal{A}\mathcal{U}|_{x_L} &= -\mathcal{V}^T Z_+ \left(\Delta_+ R_L + \widetilde{R}_L^T \Delta_- \right) Z_-^T \mathcal{U}|_{x_L} \\ \mathcal{V}^T \mathcal{A}\mathcal{U}|_{x_R} &= -\mathcal{V}^T Z_- \left(\widetilde{R}_R^T \Delta_+ + \Delta_- R_R \right) Z_+^T \mathcal{U}|_{x_R} \end{aligned}$$

and if (18) holds, then $\mathcal{V}^T \mathcal{A}\mathcal{U} = 0$ at both boundaries and the above claim is confirmed.

Remark 2.3. A functional can also include outgoing solution terms from the boundary, as $\mathcal{J}(\mathcal{U}) = \langle \mathcal{G}, \mathcal{U} \rangle + \alpha \Delta_+ Z_+^T \mathcal{U}|_{x_R} + \beta \Delta_- Z_-^T \mathcal{U}|_{x_L}$. This would specify the boundary data in (16) to $\widetilde{g}_L = -\widetilde{P}_L \beta^T$ and $\widetilde{g}_R = \widetilde{P}_R \alpha^T$, compare with [9]. For simplicity we consider $\mathcal{J}(\mathcal{U}) = \langle \mathcal{G}, \mathcal{U} \rangle$ which means that actually, the boundary data in (16) is zero.

2.4.1 Well-posedness of the dual problem

The growth rate for the dual problem is given by

$$\frac{d}{d\tau} \|\mathcal{V}\|^2 + 2\langle \mathcal{V}, \mathcal{R}\mathcal{V} \rangle = \text{BT}_L^{\text{dual}} + \text{BT}_R^{\text{dual}}$$

where the boundary terms are (the homogeneous boundary conditions have been applied)

$$\text{BT}_L^{\text{dual}} = \mathcal{V}^T Z_+ \widetilde{\mathcal{C}}_L Z_+^T \mathcal{V}|_{x_L}, \quad \text{BT}_R^{\text{dual}} = \mathcal{V}^T Z_- \widetilde{\mathcal{C}}_R Z_-^T \mathcal{V}|_{x_R}$$

and where $\widetilde{\mathcal{C}}_L = -\Delta_+ - \Delta_+ R_L \Delta_-^{-1} R_L^T \Delta_+$ and $\widetilde{\mathcal{C}}_R = \Delta_- + \Delta_- R_R \Delta_+^{-1} R_R^T \Delta_-$. For well-posedness of the dual problem $\widetilde{\mathcal{C}}_L \leq 0$ and $\widetilde{\mathcal{C}}_R \leq 0$ are necessary.

Recall that the primal problem is well-posed if $\mathcal{C}_L, \mathcal{C}_R \leq 0$. The dual demand $\widetilde{\mathcal{C}}_L \leq 0$ is directly fulfilled if $\mathcal{C}_L \leq 0$ and $\widetilde{\mathcal{C}}_R \leq 0$ follows from $\mathcal{C}_R \leq 0$. (In the special case when R_L, R_R are square, invertible matrices, this is trivial. For general R_L, R_R it can be shown with the help of Sylvester's determinant theorem.) We conclude that the dual problem (16) with (17), (18) is well-posed if the primal problem (4) with (10) is well-posed.

Remark 2.4. In [2, 3] the dual consistent schemes are constructed by first designing the boundary conditions (for incompletely parabolic problems) such that both the primal and the dual problem are well-posed. Their different approach can partly be explained by their wish to have the boundary conditions in the special form $H_{L,R}U \mp BU_x = G_{L,R}$. Looking e.g. at Eq. (30) in [2], we note that after applying the boundary conditions, $U^T M_L U \geq 0$ is needed for stability. However, if B is singular, replacing BU_x by $\pm H_{L,R}U$ does not guarantee that all conditions have been completely used, and u and p in $U = [p, u]^T$ in $U^T M_L U$ can be linearly dependent. Therefore the demand $M_L \geq 0$ in (31) is unnecessarily strong and gives some extra restrictions on the boundary conditions.

2.4.2 Discretization of the dual problem

The semi-discrete scheme approximating the dual problem (16) is written

$$\begin{aligned} V_\tau + (I_N \otimes \mathcal{R})V - (D_1 \otimes \mathcal{A})V &= G + (H^{-1}e_0 \otimes \widetilde{\Sigma}_0)(\widetilde{\mathcal{B}}_L V_0 - \widetilde{g}_L) \\ &\quad + (H^{-1}e_N \otimes \widetilde{\Sigma}_N)(\widetilde{\mathcal{B}}_R V_N - \widetilde{g}_R), \end{aligned} \quad (20)$$

where $V_i(\tau)$ represents $\mathcal{V}(x_i, \tau)$. The SAT parameters $\widetilde{\Sigma}_0$ and $\widetilde{\Sigma}_N$ are yet unknown.

2.5 Dual consistency

The semi-discrete scheme (11) is rewritten as $U_t + LU = \text{RHS}$, where

$$L = (I_N \otimes \mathcal{R}) + (D_1 \otimes \mathcal{A}) - (H^{-1}E_0 \otimes \Sigma_0 \mathcal{B}_L) - (H^{-1}E_N \otimes \Sigma_N \mathcal{B}_R)$$

and where RHS only depends on known data. In contrast to the continuous counterpart \mathcal{L} , L includes the boundary conditions explicitly. According to [2], the discrete adjoint operator is given by $L^* = \overline{H}^{-1} L^T \overline{H}$, which, using (12), leads to

$$L^* = (I_N \otimes \mathcal{R}) - (D_1 \otimes \mathcal{A}) - (H^{-1}E_0 \otimes \mathcal{B}_L^T \Sigma_0^T + \mathcal{A}) - (H^{-1}E_N \otimes \mathcal{B}_R^T \Sigma_N^T - \mathcal{A}) \quad (21)$$

If L^* is a consistent approximation of \mathcal{L}^* in (19), then the scheme (11) is dual consistent. Looking at (20), we see that L^* must have the form

$$(L^*)^{\text{goal}} = (I_N \otimes \mathcal{R}) - (D_1 \otimes \mathcal{A}) - (H^{-1}E_0 \otimes \widetilde{\Sigma}_0 \widetilde{\mathcal{B}}_L) - (H^{-1}E_N \otimes \widetilde{\Sigma}_N \widetilde{\mathcal{B}}_R). \quad (22)$$

Thus we have dual consistency if the expressions in (21) and (22) are equal. This gives us the following requirements:

$$\mathcal{B}_L^T \Sigma_0^T + \mathcal{A} - \widetilde{\Sigma}_0 \widetilde{\mathcal{B}}_L = 0 \quad \mathcal{B}_R^T \Sigma_N^T - \mathcal{A} - \widetilde{\Sigma}_N \widetilde{\mathcal{B}}_R = 0.$$

Similarly to the penalty parameters (14) for the primal problem, we make the ansätze

$$\widetilde{\Sigma}_0 = (Z_+ \widetilde{\Gamma}_0 + Z_- \widetilde{\Pi}_0) \widetilde{P}_L^{-1} \quad \widetilde{\Sigma}_N = (Z_+ \widetilde{\Pi}_N + Z_- \widetilde{\Gamma}_N) \widetilde{P}_R^{-1} \quad (23)$$

for the penalty parameters of the dual problem. We consider the left boundary and use (14) and (23), together with (7), (10) and (17), to write

$$\mathcal{B}_L^T \Sigma_0^T + \mathcal{A} - \widetilde{\Sigma}_0 \widetilde{\mathcal{B}}_L = \begin{bmatrix} Z_+^T \\ Z_-^T \end{bmatrix}^T \begin{bmatrix} \Delta_+ + \Pi_0^T - \widetilde{\Gamma}_0 \widetilde{R}_L & \Gamma_0^T - \widetilde{\Gamma}_0 \\ R_L^T \Pi_0^T - \widetilde{\Pi}_0 \widetilde{R}_L & \Delta_- + R_L^T \Gamma_0^T - \widetilde{\Pi}_0 \end{bmatrix} \begin{bmatrix} Z_+^T \\ Z_-^T \end{bmatrix}$$

which is zero if and only if the four entries of the matrix are zero. These four demands are rearranged to the more convenient form

$$\Pi_0 = -\Delta_+ - \Delta_+ R_L \Delta_-^{-1} \Gamma_0 \quad (24a)$$

$$\widetilde{R}_L = -\Delta_-^{-1} R_L^T \Delta_+ \quad (24b)$$

$$\widetilde{\Gamma}_0 = \Gamma_0^T \quad (24c)$$

$$\widetilde{\Pi}_0 = \Delta_- - \Delta_- \widetilde{R}_L \Delta_+^{-1} \widetilde{\Gamma}_0. \quad (24d)$$

Note that (24a) only depends on parameters from the primal problem, while (24d) only depends on parameters from the dual problem. Interestingly enough, (24b) is nothing

but the duality demand (18) for the continuous problem. The demand (24c) relates the penalty of the dual problem to the primal penalty.

Unless we actually want to solve the dual problem, it is enough to consider the first demand, (24a). We repeat the above derivation also for the right boundary and get the following result: The penalty parameters Σ_0 and Σ_N in (14) with

$$\Pi_0 = -\Delta_+ - \Delta_+ R_L \Delta_-^{-1} \Gamma_0, \quad \Pi_N = \Delta_- - \Delta_- R_R \Delta_+^{-1} \Gamma_N, \quad (25)$$

makes the discretization (11) dual consistent.

Remark 2.5. If the discrete primal problem (11) is dual consistent there is no need to check if the discrete dual problem (20) is stable – in [8] it is stated that stability of the primal problem implies stability of the dual problem, because the system matrix for the dual problem is the transpose of the system matrix for the primal problem – that is the primal and dual discrete problems have exactly the same growth rates for zero data.

2.6 Penalty parameters for the hyperbolic problem

Consider the penalty parameter ansatz for the left boundary, $\Sigma_0 = (Z_+ \Pi_0 + Z_- \Gamma_0) P_L^{-1}$, which is given in (14). From a stability point of view, we must choose Π_0 and Γ_0 such that C_0 in (15) becomes non-positive. In addition, for dual consistency the constraint in (25) must be fulfilled. By inserting the duality constraint $\Pi_0 = -\Delta_+ - \Delta_+ R_L \Delta_-^{-1} \Gamma_0$ from (25) into C_0 we obtain, after some rearrangements, the expression

$$C_0 = \begin{bmatrix} P_L^{-1} \mathcal{B}_L \\ Z_-^T \end{bmatrix}^T \begin{bmatrix} -\Delta_+ - \Delta_+ R_L \Delta_-^{-1} \Gamma_0 - (\Delta_+ R_L \Delta_-^{-1} \Gamma_0)^T & \Gamma_0^T \Delta_-^{-1} \mathcal{C}_L \\ \mathcal{C}_L \Delta_-^{-1} \Gamma_0 & \mathcal{C}_L \end{bmatrix} \begin{bmatrix} P_L^{-1} \mathcal{B}_L \\ Z_-^T \end{bmatrix}.$$

The most obvious strategy to make $C_0 \leq 0$ is to cancel the off-diagonal entries by putting $\Gamma_0 = 0$, but note that other choices exist. To single out the optimal (in a certain sense) candidate, we use another approach. With (7), (10) and $\tilde{g}_L = P_L^{-1} g_L$, the left boundary term in (13) can be rearranged as

$$\begin{aligned} \text{BT}_L^{\text{Disc.}} &= U_0^T Z_- (\Delta_- + R_L^T \Delta_+ R_L) Z_-^T U_0 - 2\tilde{g}_L^T \Delta_+ R_L Z_-^T U_0 + \tilde{g}_L^T \Delta_+ \tilde{g}_L \\ &\quad - (\mathcal{B}_L U_0 - g_L)^T P_L^{-T} \Delta_+ P_L^{-1} (\mathcal{B}_L U_0 - g_L) \\ &\quad + 2(\mathcal{B}_L U_0 - g_L)^T (\Sigma_0 + Z_+ \Delta_+ P_L^{-1})^T U_0 \end{aligned} \quad (26)$$

where we see that the first row corresponds exactly to the continuous boundary term BT_L in (9). The second row is a damping term that is quadratically proportional to the solution's deviation from data at the boundary, $\mathcal{B}_L U_0 - g_L$. The term in the last row is only linearly proportional to this deviation, so we would prefer it to be zero. This is possible if the penalty parameter is chosen exactly as $\Sigma_0 = -Z_+ \Delta_+ P_L^{-1}$. Luckily this choice fulfills both the stability requirement and the duality constraint. We repeat the above derivation also for the right boundary and summarize our findings in Theorem 2.6.

Theorem 2.6. *Consider the problem (4) with an associated factorization (6) where Z is non-singular. With the particular choice of penalty parameters*

$$\Sigma_0 = -Z_+ \Delta_+ P_L^{-1}, \quad \Sigma_N = Z_- \Delta_- P_R^{-1}, \quad (27)$$

the scheme (11) is a stable and dual consistent discretization of (4). The matrices P_L and P_R are specified through (10).

Proof. Comparing with (14), we note that Σ_0 in (27) is obtained using $\Pi_0 = -\Delta_+$ and $\Gamma_0 = 0$. These values fulfill the left duality constraint in (25). Inserting $\Gamma_0 = 0$ into C_0 above, we obtain $C_0 = Z_- \mathcal{C}_L Z_-^T - \mathcal{B}_L^T P_L^{-T} \Delta_+ P_L^{-1} \mathcal{B}_L$, which is negative semi-definite if the continuous problem is well-posed (in the $\mathcal{C}_L \leq 0$ sense). Thus the stability demand $C_0 \leq 0$ is fulfilled. The same is done for the right boundary, completing the proof. \square

Remark 2.7. The seemingly very specific choice of penalty parameters in Theorem 2.6 is, in fact, a family of penalty parameters, depending on the factorization used. Note that it is not necessary to use the same factorization for the left and the right boundary.

Remark 2.8. If characteristic boundary conditions (in the sense $R_L, R_R = 0$) are used, the scheme (11) together with the SATs from Theorem 2.6 simplifies to

$$U_t + (I_N \otimes \mathcal{R})U + (D_1 \otimes \mathcal{A})U = F + (H^{-1}E_0 \otimes -\mathcal{A}_+)U + (H^{-1}E_N \otimes \mathcal{A}_-)U$$

in the homogeneous case, where $\mathcal{A}_+ = Z_+ \Delta_+ Z_+^T$ and $\mathcal{A}_- = Z_- \Delta_- Z_-^T$. When the factorization refers to the eigendecomposition, this corresponds to the SAT used for the characteristic boundary conditions of the nonlinear Euler equations in [9].

3 Parabolic systems

Consider the parabolic (or incompletely parabolic) system of partial differential equations

$$\begin{aligned} \mathcal{U}_t + \mathcal{A}\mathcal{U}_x - \mathcal{E}\mathcal{U}_{xx} &= \mathcal{F}, & x \in [x_L, x_R], \\ \mathcal{H}_L\mathcal{U} + \mathcal{G}_L\mathcal{U}_x &= g_L, & x = x_L, \\ \mathcal{H}_R\mathcal{U} + \mathcal{G}_R\mathcal{U}_x &= g_R, & x = x_R, \end{aligned} \quad (28)$$

for $t \geq 0$, augmented with the initial condition $\mathcal{U}(x, 0) = \mathcal{U}_0(x)$. The matrices \mathcal{A} and $\mathcal{E} \geq 0$ are symmetric $n \times n$ matrices, and we assume that \mathcal{G}_L and \mathcal{G}_R scales as $\mathcal{G}_L = \mathcal{K}_L \mathcal{E}$ and $\mathcal{G}_R = \mathcal{K}_R \mathcal{E}$, respectively. Treating \mathcal{U}_x as a separate variable, we can rewrite (28) as a first order system (as was also done in [9, 1]), arriving at

$$\begin{aligned} \overline{\mathcal{I}}\overline{\mathcal{U}}_t + \overline{\mathcal{R}}\overline{\mathcal{U}} + \overline{\mathcal{A}}\overline{\mathcal{U}}_x &= \overline{\mathcal{F}}, & x \in [x_L, x_R], \\ \overline{\mathcal{B}}_L\overline{\mathcal{U}} &= g_L, & x = x_L, \\ \overline{\mathcal{B}}_R\overline{\mathcal{U}} &= g_R, & x = x_R, \end{aligned} \quad (29)$$

where

$$\bar{\mathcal{I}} = \begin{bmatrix} I_n & 0 \\ 0 & 0 \end{bmatrix}, \quad \bar{\mathcal{R}} = \begin{bmatrix} 0 & 0 \\ 0 & \mathcal{E} \end{bmatrix}, \quad \bar{\mathcal{U}} = \begin{bmatrix} \mathcal{U} \\ \mathcal{U}_x \end{bmatrix}, \quad \bar{\mathcal{F}} = \begin{bmatrix} \mathcal{F} \\ 0 \end{bmatrix}$$

and

$$\bar{\mathcal{A}} = \begin{bmatrix} \mathcal{A} & -\mathcal{E} \\ -\mathcal{E} & 0 \end{bmatrix}, \quad \bar{\mathcal{B}}_L = [\mathcal{H}_L \quad \mathcal{G}_L], \quad \bar{\mathcal{B}}_R = [\mathcal{H}_R \quad \mathcal{G}_R]. \quad (30)$$

The system (29) has almost the same form as (4) since $\bar{\mathcal{R}} \geq 0$ and $\bar{\mathcal{A}}$ are symmetric $m \times m$ matrices, where $m = 2n$. Thus we can use the results from the hyperbolic case.

Remark 3.1. In [2, 3] the operators corresponding to \mathcal{H}_L , \mathcal{G}_L , \mathcal{H}_R and \mathcal{G}_R are square $n \times n$ matrices and their ranks are changed to suit the number of boundary conditions. We adapt the matrix dimensions instead. Both approaches have their respective advantages.

3.1 Discretization using wide-stencil second derivative operators

To discretize the parabolic problem, we first consider the reformulated problem (29), and use the results from the hyperbolic section. Then we rearrange the terms such that we get an equivalent scheme but in a form corresponding to (28). These steps, which are done in Appendix A, lead to

$$U_t + (D_1 \otimes \mathcal{A})U - (D_1^2 \otimes \mathcal{E})U = F + \bar{H}^{-1}(e_0 \otimes \hat{\mu}_0 + D_1^T e_0 \otimes \hat{\nu}_0)\hat{\xi}_0 + \bar{H}^{-1}(e_N \otimes \hat{\mu}_N + D_1^T e_N \otimes \hat{\nu}_N)\hat{\xi}_N \quad (31)$$

where

$$\hat{\xi}_0 = \mathcal{H}_L U_0 + \mathcal{G}_L(\bar{D}U)_0 - g_L, \quad \hat{\xi}_N = \mathcal{H}_R U_N + \mathcal{G}_R(\bar{D}U)_N - g_R, \quad (32)$$

and $\bar{H} = (H \otimes I_n)$ and $\bar{D} = (D_1 \otimes I_n)$. The penalty parameters in (31) are

$$\begin{aligned} \hat{\mu}_0 &= (-\bar{Z}_1 + \hat{q}\bar{Z}_2)\bar{\Delta}_+(\bar{P}_L + \hat{q}\mathcal{K}_L\bar{Z}_2\bar{\Delta}_+)^{-1}, & \hat{\nu}_0 &= \bar{Z}_2\bar{\Delta}_+(\bar{P}_L + \hat{q}\mathcal{K}_L\bar{Z}_2\bar{\Delta}_+)^{-1} \\ \hat{\mu}_N &= (\bar{Z}_3 + \hat{q}\bar{Z}_4)\bar{\Delta}_-(\bar{P}_R - \hat{q}\mathcal{K}_R\bar{Z}_4\bar{\Delta}_-)^{-1}, & \hat{\nu}_N &= -\bar{Z}_4\bar{\Delta}_-(\bar{P}_R - \hat{q}\mathcal{K}_R\bar{Z}_4\bar{\Delta}_-)^{-1} \end{aligned} \quad (33)$$

where the matrices $\bar{Z}_{1,2,3,4}$ are defined through

$$\bar{Z}_+ = \begin{bmatrix} \bar{Z}_1 \\ \bar{Z}_2 \end{bmatrix}, \quad \bar{Z}_- = \begin{bmatrix} \bar{Z}_3 \\ \bar{Z}_4 \end{bmatrix}. \quad (34)$$

As before, $\bar{\Delta}_\pm$, \bar{Z}_\pm and \bar{P}_L , \bar{P}_R are described in (6) and (10), respectively, but are now obtained using $\bar{\mathcal{A}}$ and $\bar{\mathcal{B}}_L$, $\bar{\mathcal{B}}_R$ from (30). Finally, the quantity \hat{q} in (33) is given by

$$\hat{q} = e_0^T H^{-1} e_0 = e_N^T H^{-1} e_N. \quad (35)$$

The matrix H is positive definite and proportional to the grid size h , and thus \hat{q} is a positive scalar proportional to $1/h$.

3.2 Discretization using narrow-stencil second derivative operators

In [1], it was suggested that dual consistency might require wide-stencil second derivative operators, but next we will show that this is not necessary. The semi-discrete scheme approximating (28) is now written, analogously to (31), as

$$U_t + (D_1 \otimes \mathcal{A})U - (D_2 \otimes \mathcal{E})U = F + \bar{H}^{-1}(e_0 \otimes \mu_0 + S^T e_0 \otimes \nu_0)\xi_0 \\ + \bar{H}^{-1}(e_N \otimes \mu_N + S^T e_N \otimes \nu_N)\xi_N. \quad (36)$$

The operator D_2 , which approximates the second derivative operator, is no longer limited to the previous form D_1^2 , where the first derivative is used twice. However, D_2 must still fulfill the SBP relations

$$D_2 = H^{-1}(-A_S + (E_N - E_0)S), \quad A_S = A_S^T = S^T M S \geq 0. \quad (37)$$

The first and last row of the matrix S are consistent difference stencils, see e.g. [13]. For dual consistency, A_S must be symmetric. Further, we have

$$\xi_0 = \mathcal{H}_L U_0 + \mathcal{G}_L(\bar{S}U)_0 - g_L, \quad \xi_N = \mathcal{H}_R U_N + \mathcal{G}_R(\bar{S}U)_N - g_R, \quad (38)$$

where

$$\bar{S} = S \otimes I_n, \quad (\bar{S}U)_0 = (e_0^T S \otimes I_n)U, \quad (\bar{S}U)_N = (e_N^T S \otimes I_n)U.$$

We also define

$$q \equiv q_0 + |q_c| = q_N + |q_c| \quad (39)$$

where

$$q_0 = e_0^T M^{-1} e_0, \quad q_N = e_N^T M^{-1} e_N, \quad q_c = e_0^T M^{-1} e_N = e_N^T M^{-1} e_0, \quad (40)$$

and where M is a part of D_2 as stated in (37). In Section 4 we provide q for various D_2 matrices. The penalty parameters μ_0 , ν_0 , μ_N and ν_N in (36) are now given by:

Theorem 3.2. *Consider the problem (28) with $\mathcal{G}_L = \mathcal{K}_L \mathcal{E}$ and $\mathcal{G}_R = \mathcal{K}_R \mathcal{E}$. Further, let $\bar{\mathcal{A}}$, which is specified in (30), be factorized as $\bar{\mathcal{A}} = \bar{Z} \bar{\Delta} \bar{Z}^T$ as described in (6). Then the particular choice of penalty parameters*

$$\mu_0 = (-\bar{Z}_1 + q\bar{Z}_2)\bar{\Delta}_+(\bar{P}_L + q\mathcal{K}_L\bar{Z}_2\bar{\Delta}_+)^{-1}, \quad \nu_0 = \bar{Z}_2\bar{\Delta}_+(\bar{P}_L + q\mathcal{K}_L\bar{Z}_2\bar{\Delta}_+)^{-1} \\ \mu_N = (\bar{Z}_3 + q\bar{Z}_4)\bar{\Delta}_-(\bar{P}_R - q\mathcal{K}_R\bar{Z}_4\bar{\Delta}_-)^{-1}, \quad \nu_N = -\bar{Z}_4\bar{\Delta}_-(\bar{P}_R - q\mathcal{K}_R\bar{Z}_4\bar{\Delta}_-)^{-1} \quad (41)$$

makes the scheme in (36) stable and dual consistent. The matrices $\bar{Z}_{1,2,3,4}$ are given in (34), \bar{P}_L , \bar{P}_R are obtained from (10) (using $\bar{\mathcal{B}}_L$, $\bar{\mathcal{B}}_R$ in (30)) and q is defined in (39).

Note that q in (39) is a generalization of \widehat{q} in (35), and that the penalty parameters in (41) and (33) are identical if $q = \widehat{q}$. Hence the narrow-stencil scheme (36) is a generalization of the wide-stencil scheme in (31), since the schemes are identical if we choose $D_2 = D_1^2$, $S = D_1$ and $M = H$. In the rest of this section we will justify these generalizations and prove Theorem 3.2 by showing that the penalties given in (41) indeed make the scheme (36) stable and dual consistent.

3.3 Stability when using narrow-stencil second derivative operators

We multiply the scheme (36) by $U^T \overline{H}$ from the left and add the transpose of the result. Thereafter using the SBP-properties in (12) and (37) yields

$$\frac{d}{dt} \|U\|_H^2 + 2U^T (S^T M S \otimes \mathcal{E}) U = 2\langle U, F \rangle_H + \text{BT}_L^{Disc.} + \text{BT}_R^{Disc.}, \quad (42)$$

where

$$\begin{aligned} \text{BT}_L^{Disc.} &= U_0^T \mathcal{A} U_0 - 2U_0^T \mathcal{E} (\overline{S} U)_0 + 2(U_0^T \mu_0 + (\overline{S} U)_0^T \nu_0) \xi_0 \\ \text{BT}_R^{Disc.} &= -U_N^T \mathcal{A} U_N + 2U_N^T \mathcal{E} (\overline{S} U)_N + 2(U_N^T \mu_N + (\overline{S} U)_N^T \nu_N) \xi_N \end{aligned} \quad (43)$$

where $\xi_{0,N}$ are given in (38). If $\text{BT}_L^{Disc.}$ and $\text{BT}_R^{Disc.}$ are non-positive for zero data the scheme is stable. This can be achieved if μ_0 , ν_0 , μ_N and ν_N are chosen freely, but the scheme should also be dual consistent. It turns out that in some cases these requirements are impossible to combine, for example when having Dirichlet boundary conditions. We therefore need an alternative way to show stability.

First, we assume that the penalty parameters ν_0 and ν_N scales with \mathcal{E} . Let

$$\nu_0 = -\mathcal{E} \kappa_0, \quad \nu_N = -\mathcal{E} \kappa_N. \quad (44)$$

Next, we take a look at the wide case (which is partly presented in Appendix A). Using a wide counterpart to (44), $\widehat{\nu}_0 = -\mathcal{E} \widehat{\kappa}_0$ and $\widehat{\nu}_N = -\mathcal{E} \widehat{\kappa}_N$, and the later relations in (71) and (72), we can rewrite (67b) as

$$\widehat{W} = \overline{D} U + (H^{-1} e_0 \otimes \widehat{\kappa}_0) \widehat{\xi}_0 + (H^{-1} e_N \otimes \widehat{\kappa}_N) \widehat{\xi}_N.$$

We return to the narrow-stencil scheme (36). Inspired by the wide case, we define

$$W \equiv \overline{S} U + (M^{-1} e_0 \otimes \kappa_0) \xi_0 + (M^{-1} e_N \otimes \kappa_N) \xi_N. \quad (45)$$

From (45) we compute

$$\begin{aligned} W^T (M \otimes \mathcal{E}) W &= U^T (S^T M S \otimes \mathcal{E}) U + (2(\overline{S} U)_0 + q_0 \kappa_0 \xi_0 + q_c \kappa_N \xi_N)^T \mathcal{E} \kappa_0 \xi_0 \\ &\quad + (2(\overline{S} U)_N + q_N \kappa_N \xi_N + q_c \kappa_0 \xi_0)^T \mathcal{E} \kappa_N \xi_N \end{aligned}$$

where q_0 , q_N and q_c are given in (40). In the general case, q_c can be non-zero. Since we want to treat the two boundaries separately, we use Young's inequality, $q_c(\xi_N^T \kappa_N^T \mathcal{E} \kappa_0 \xi_0 + \xi_0^T \kappa_0^T \mathcal{E} \kappa_N \xi_N) \leq |q_c| (\xi_0^T \kappa_0^T \mathcal{E} \kappa_0 \xi_0 + \xi_N^T \kappa_N^T \mathcal{E} \kappa_N \xi_N)$, which leads to

$$W^T(M \otimes \mathcal{E})W \leq U^T(S^T M S \otimes \mathcal{E})U + (2(\bar{S}U)_0 + q\kappa_0 \xi_0)^T \mathcal{E} \kappa_0 \xi_0 + (2(\bar{S}U)_N + q\kappa_N \xi_N)^T \mathcal{E} \kappa_N \xi_N \quad (46)$$

where $q = q_0 + |q_c| = q_N + |q_c|$, as stated in (39). Further, we note that multiplying (45) by $(e_0^T \otimes I_n)$ and $(e_N^T \otimes I_n)$, respectively, yields the relations $W_0 = (\bar{S}U)_0 + q_0 \kappa_0 \xi_0 + q_c \kappa_N \xi_N$ and $W_N = (\bar{S}U)_N + q_c \kappa_0 \xi_0 + q_N \kappa_N \xi_N$. Instead of using those, which contain unwanted terms from the other boundary, we define

$$\widetilde{W}_0 \equiv (\bar{S}U)_0 + q\kappa_0 \xi_0 \quad \widetilde{W}_N \equiv (\bar{S}U)_N + q\kappa_N \xi_N. \quad (47)$$

Inserting the relation (46) into (42), we obtain

$$\frac{d}{dt} \|U\|_H^2 + 2W^T(M \otimes \mathcal{E})W \leq 2\langle U, F \rangle_H + \widetilde{\text{BT}}_L^{Disc.} + \widetilde{\text{BT}}_R^{Disc.} \quad (48)$$

where (43) and (47) together with (44) yields

$$\begin{aligned} \widetilde{\text{BT}}_L^{Disc.} &= U_0^T \mathcal{A} U_0 - 2U_0^T \mathcal{E} \widetilde{W}_0 + 2(U_0^T (\mu_0 - q\nu_0) - \widetilde{W}_0^T \nu_0) \xi_0 \\ \widetilde{\text{BT}}_R^{Disc.} &= -U_N^T \mathcal{A} U_N + 2U_N^T \mathcal{E} \widetilde{W}_N + 2(U_N^T (\mu_N + q\nu_N) - \widetilde{W}_N^T \nu_N) \xi_N. \end{aligned} \quad (49)$$

If the penalty parameters make $\widetilde{\text{BT}}_L^{Disc.} \leq 0$ and $\widetilde{\text{BT}}_R^{Disc.} \leq 0$ for zero data, (36) is stable.

Again taking the left boundary as an example, we define $\widetilde{U}_0 = [U_0^T, \widetilde{W}_0^T]^T$ and write the first part of $\widetilde{\text{BT}}_L^{Disc.}$ in (49) as

$$U_0^T \mathcal{A} U_0 - 2U_0^T \mathcal{E} \widetilde{W}_0 = \widetilde{U}_0^T \widetilde{\mathcal{A}} \widetilde{U}_0. \quad (50)$$

Next, using the relations (30), (47) and (38), recalling the assumptions $\mathcal{G}_L = \mathcal{K}_L \mathcal{E}$ and $\nu_0 = -\mathcal{E} \kappa_0$, and thereafter using (41) from Theorem 3.2, we obtain

$$\bar{\mathcal{B}}_L \widetilde{U}_0 - g_L = \bar{P}_L (\bar{P}_L + q \mathcal{K}_L \bar{\mathcal{Z}}_2 \bar{\Delta}_+)^{-1} \xi_0. \quad (51)$$

From (41) we also get

$$\mu_0 - q\nu_0 = -\bar{\mathcal{Z}}_1 \bar{\Delta}_+ (\bar{P}_L + q \mathcal{K}_L \bar{\mathcal{Z}}_2 \bar{\Delta}_+)^{-1}, \quad -\nu_0 = -\bar{\mathcal{Z}}_2 \bar{\Delta}_+ (\bar{P}_L + q \mathcal{K}_L \bar{\mathcal{Z}}_2 \bar{\Delta}_+)^{-1}$$

such that the second part of $\widetilde{\text{BT}}_L^{Disc.}$ in (49) becomes

$$2 \left(U_0^T (\mu_0 - q\nu_0) - \widetilde{W}_0^T \nu_0 \right) \xi_0 = 2\widetilde{U}_0^T \bar{\Sigma}_0 (\bar{\mathcal{B}}_L \widetilde{U}_0 - g_L) \quad (52)$$

where the relations (34) and (51) have been used, and where $\bar{\Sigma}_0 = -\bar{Z}_+ \bar{\Delta}_+ \bar{P}_L^{-1}$. Now we can, by inserting (50) and (52) into (49), write

$$\widetilde{\text{BT}}_L^{Disc.} = \tilde{U}_0^T \bar{\mathcal{A}} \tilde{U}_0 + 2\tilde{U}_0^T \bar{\Sigma}_0 (\bar{\mathcal{B}}_L \tilde{U}_0 - g_L)$$

which has exactly the same form as $\text{BT}_L^{Disc.}$ in (13). We thus know that $\widetilde{\text{BT}}_L^{Disc.} \leq 0$ for zero data, since $\bar{\Sigma}_0$ is computed just as in the hyperbolic case. The same procedure can, of course, be repeated for the right boundary. We conclude that the scheme (36) with the penalty parameters (41) is stable.

3.4 Dual consistency for narrow-stencil second derivative operators

The dual problem of (28) is

$$\begin{aligned} \mathcal{V}_\tau - \mathcal{A}\mathcal{V}_x - \mathcal{E}\mathcal{V}_{xx} &= \mathcal{G}, & x \in [x_L, x_R], \\ \widetilde{\mathcal{H}}_L \mathcal{V} + \widetilde{\mathcal{G}}_L \mathcal{V}_x &= \widetilde{g}_L, & x = x_L, \\ \widetilde{\mathcal{H}}_R \mathcal{V} + \widetilde{\mathcal{G}}_R \mathcal{V}_x &= \widetilde{g}_R, & x = x_R, \end{aligned} \quad (53)$$

for $\tau \geq 0$ and with $\mathcal{V}(x, 0) = \mathcal{V}_0(x)$. The spatial operator in (28) and its dual are thus

$$\mathcal{L} = \mathcal{A} \frac{\partial}{\partial x} - \mathcal{E} \frac{\partial^2}{\partial x^2}, \quad \mathcal{L}^* = -\mathcal{A} \frac{\partial}{\partial x} - \mathcal{E} \frac{\partial^2}{\partial x^2}. \quad (54)$$

The semi-discrete approximation of (53) is

$$\begin{aligned} V_\tau - (D_1 \otimes \mathcal{A})V - (D_2 \otimes \mathcal{E})V &= G + \bar{H}^{-1}(e_0 \otimes \widetilde{\mu}_0 + S^T e_0 \otimes \widetilde{\nu}_0) \widetilde{\xi}_0 \\ &\quad + \bar{H}^{-1}(e_N \otimes \widetilde{\mu}_N + S^T e_N \otimes \widetilde{\nu}_N) \widetilde{\xi}_N, \end{aligned} \quad (55)$$

where

$$\widetilde{\xi}_0 = \widetilde{\mathcal{H}}_L V_0 + \widetilde{\mathcal{G}}_L (\bar{S}V)_0 - \widetilde{g}_L, \quad \widetilde{\xi}_N = \widetilde{\mathcal{H}}_R V_N + \widetilde{\mathcal{G}}_R (\bar{S}V)_N - \widetilde{g}_R.$$

From (36) we see that the discrete operator, corresponding to \mathcal{L} in (54), is

$$\begin{aligned} L &= (D_1 \otimes \mathcal{A}) - (D_2 \otimes \mathcal{E}) - \bar{H}^{-1}(e_0 \otimes \mu_0 + S^T e_0 \otimes \nu_0)(e_0^T \otimes \mathcal{H}_L + e_0^T S \otimes \mathcal{G}_L) \\ &\quad - \bar{H}^{-1}(e_N \otimes \mu_N + S^T e_N \otimes \nu_N)(e_N^T \otimes \mathcal{H}_R + e_N^T S \otimes \mathcal{G}_R). \end{aligned} \quad (56)$$

Using the relations in (12) and (37), we obtain

$$\begin{aligned} L^* &= \bar{H}^{-1} L^T \bar{H} = -(D_1 \otimes \mathcal{A}) - (D_2 \otimes \mathcal{E}) \\ &\quad - \bar{H}^{-1}(e_0 e_0^T \otimes \mathcal{A}) + \bar{H}^{-1}((S^T e_0 e_0^T - e_0 e_0^T S) \otimes \mathcal{E}) \\ &\quad + \bar{H}^{-1}(e_N e_N^T \otimes \mathcal{A}) - \bar{H}^{-1}((S^T e_N e_N^T - e_N e_N^T S) \otimes \mathcal{E}) \\ &\quad - \bar{H}^{-1}(e_0 \otimes \mathcal{H}_L^T + S^T e_0 \otimes \mathcal{G}_L^T)(e_0^T \otimes \mu_0^T + e_0^T S \otimes \nu_0^T) \\ &\quad - \bar{H}^{-1}(e_N \otimes \mathcal{H}_R^T + S^T e_N \otimes \mathcal{G}_R^T)(e_N^T \otimes \mu_N^T + e_N^T S \otimes \nu_N^T). \end{aligned}$$

However, from (55) we see that for dual consistency L^* must have the form

$$\begin{aligned} (L^*)^{goal} &= -(D_1 \otimes \mathcal{A}) - (D_2 \otimes \mathcal{E}) \\ &\quad - \bar{H}^{-1}(e_0 \otimes \widetilde{\mu}_0 + S^T e_0 \otimes \widetilde{\nu}_0)(e_0^T \otimes \widetilde{\mathcal{H}}_L + e_0^T S \otimes \widetilde{\mathcal{G}}_L) \\ &\quad - \bar{H}^{-1}(e_N \otimes \widetilde{\mu}_N + S^T e_N \otimes \widetilde{\nu}_N)(e_N^T \otimes \widetilde{\mathcal{H}}_R + e_N^T S \otimes \widetilde{\mathcal{G}}_R). \end{aligned}$$

Demanding that $L^* = (L^*)^{goal}$, gives us the duality constraints

$$\begin{aligned} \begin{bmatrix} \mathcal{H}_L^T \mu_0^T + \mathcal{A} & \mathcal{H}_L^T \nu_0^T + \mathcal{E} \\ \mathcal{G}_L^T \mu_0^T - \mathcal{E} & \mathcal{G}_L^T \nu_0^T \end{bmatrix} &= \begin{bmatrix} \widetilde{\mu}_0 \widetilde{\mathcal{H}}_L & \widetilde{\mu}_0 \widetilde{\mathcal{G}}_L \\ \widetilde{\nu}_0 \widetilde{\mathcal{H}}_L & \widetilde{\nu}_0 \widetilde{\mathcal{G}}_L \end{bmatrix} \\ \begin{bmatrix} \mathcal{H}_R^T \mu_N^T - \mathcal{A} & \mathcal{H}_R^T \nu_N^T - \mathcal{E} \\ \mathcal{G}_R^T \mu_N^T + \mathcal{E} & \mathcal{G}_R^T \nu_N^T \end{bmatrix} &= \begin{bmatrix} \widetilde{\mu}_N \widetilde{\mathcal{H}}_R & \widetilde{\mu}_N \widetilde{\mathcal{G}}_R \\ \widetilde{\nu}_N \widetilde{\mathcal{H}}_R & \widetilde{\nu}_N \widetilde{\mathcal{G}}_R \end{bmatrix}. \end{aligned} \tag{57}$$

The duality constraints in (57) do not depend explicitly on the grid size h . Moreover, we already know that for the wide case, the penalty parameters in (33) – even though they contain the h -dependent constant \hat{q} – gives dual consistency. Since the generalized penalty parameters in (41) have exactly the same form (the only difference is that they depend on *another* h -dependent constant, q) they will also yield dual consistency. We have thus shown that the penalty parameters in Theorem 3.2 indeed makes the scheme (36) stable and dually consistent.

Remark 3.3. The SAT parameters in Theorem 3.2 are probably a subset of all parameters giving stability and dual consistency since the duality constraint (57) could be used in combination with some other stability proof than the one presented here.

4 Computing q

We want to compute $q = q_0 + |q_c| = q_N + |q_c|$ as stated in (39) and are thus looking for q_0 , q_N and q_c specified in (40). For wide second derivative operators, M is equal to H , and is thus well-defined. When using narrow second derivative operators, M is defined in (37) through $A_S = S^T M S$. However, only the first and last row of S are clearly specified. In for example [4, 13, 5], the interior of S is the identity matrix, and S is then invertible. A_S is singular (since $A_S = (E_N - E_0)S - H D_2$, where D_2 and the first and last row of S are consistent difference operators) and thus an invertible S implies that M is singular.

If M and S are defined such that M is singular and S not, which is often the case, we use the following strategy to find q : The relation $A_S = S^T M S$ leads to $M^{-1} = S A_S^{-1} S^T$, but since A_S is singular we define the perturbed matrix $\tilde{A}_S \equiv A_S + \delta E_0$ and compute $\tilde{M}^{-1} = S \tilde{A}_S^{-1} S^T$ instead. This is motivated by the following proposition:

Proposition 4.1. Define $\tilde{A}_S \equiv A_S + \delta E_j$, where E_j is an all-zero matrix except for the element $(E_j)_{j,j} = 1$, with $0 \leq j \leq N$. The inverse of \tilde{A}_S is $\tilde{A}_S^{-1} = J/\delta + K_j$ where J is an all-ones matrix and K_j is a matrix that does not depend on the scalar δ . A consequence of this structure is that the corners of $\tilde{M}^{-1} = S\tilde{A}_S^{-1}S^T$ are independent of δ , such that

$$q_0 = e_0^T \tilde{M}^{-1} e_0, \quad q_N = e_N^T \tilde{M}^{-1} e_N, \quad q_c = e_0^T \tilde{M}^{-1} e_N = e_N^T \tilde{M}^{-1} e_0. \quad (58)$$

Proposition 4.1 is motivated in Appendix B. In Table 1 below we provide the value of q for all second derivative operators considered in this paper. The wide-stencil operators are given by $D_2 = D_1^2$, where D_1 has the order of accuracy (2,1), (4,2), (6,3) or (8,4), paired as (interior order, boundary order). For these operators, the q values are obtained directly from the matrix H . For the narrow-stencil operators, the q values are computed according to Proposition 4.1. All examples in Table 1, except the narrow (2,0) order operator, refers to operators given in [13].

Order	Type	qh	Comment
2,0	wide	2	
4,1	wide	$\frac{48}{17} \approx 2.8235$	
6,2	wide	$\frac{43200}{13649} \approx 3.1651$	
8,3	wide	$\frac{5080320}{1498139} \approx 3.3911$	
2,0	narrow	1	See Eq. (73)
2,1	narrow	2.5	
4,2	narrow	3.986391480987749	($N = 8$)
6,3	narrow	5.322804652661742	($N = 12$)
8,4	narrow	633.69326893357	($N = 16$)

Table 1: The q -values (scaled with h) for various second derivative operators.

Remark 4.2. The SBP operators with interior order 6 and higher have free parameters, and if those parameters are chosen differently than in [13], that will affect q .

Remark 4.3. The quantity q has nothing to do with dual consistency, but indicates how the penalty should be chosen to give energy stability. As an example, consider solving the scalar problem presented below in (60) with Dirichlet boundary conditions, using the scheme (62). Using the same technique as in Section 3.3, we find that the stability demands for the (left) penalty parameter μ_0 , in three special cases of ν_0 , are

Dual consistent (see Eq. (64))	$\nu_0 = -\varepsilon$	$\mu_0 \leq -a/2 - \varepsilon q$
Method 1 (dual inconsistent)	$\nu_0 = 0$	$\mu_0 \leq -a/2 - \varepsilon q/4$
Method 2 (dual inconsistent)	$\nu_0 = \varepsilon$	$\mu_0 \leq -a/2.$

The two latter approaches are frequently used but they do not yield dual consistency.

5 Examples and numerical experiments

In this section, we give a few concrete examples of the derived penalty parameters and perform some numerical simulations. We demonstrate that these penalty parameters give superconvergent functional output not only for the wide second derivative operators but also for the narrow ones. The following procedure is used:

- i) Consider a continuous problem on the form (28), where $\mathcal{G}_L = \mathcal{K}_L \mathcal{E}$ and $\mathcal{G}_R = \mathcal{K}_R \mathcal{E}$ are required. Identify $\bar{\mathcal{A}}$ and $\bar{\mathcal{B}}_L, \bar{\mathcal{B}}_R$ according to (30).
- ii) Factorize $\bar{\mathcal{A}}$ as $\bar{\mathcal{A}} = \bar{\mathcal{Z}} \bar{\Delta} \bar{\mathcal{Z}}^T$, according to (6), where $\bar{\mathcal{Z}}$ must be non-singular.
- iii) Compute \bar{P}_L and \bar{P}_R . From (10) we see that \bar{P}_L is the first $m_+ \times m_+$ part of $\bar{\mathcal{B}}_L \bar{\mathcal{Z}}^{-T}$, and correspondingly, that \bar{P}_R is the last $m_- \times m_-$ part of $\bar{\mathcal{B}}_R \bar{\mathcal{Z}}^{-T}$, as

$$\bar{\mathcal{B}}_L \bar{\mathcal{Z}}^{-T} = \begin{bmatrix} \bar{P}_L & 0_{m_+, m_0} & \bar{P}_L \bar{R}_L \end{bmatrix}, \quad \bar{\mathcal{B}}_R \bar{\mathcal{Z}}^{-T} = \begin{bmatrix} \bar{P}_R \bar{R}_R & 0_{m_-, m_0} & \bar{P}_R \end{bmatrix}. \quad (59)$$

- iv) The problem (28) is discretized in space using the scheme (36). Rearranging the terms in the scheme yields $U_t + LU = \text{RHS}$, where L is given in (56), and where

$$\text{RHS} = F - \bar{H}^{-1}(e_0 \otimes \mu_0 + S^T e_0 \otimes \nu_0)g_L - \bar{H}^{-1}(e_N \otimes \mu_N + S^T e_N \otimes \nu_N)g_R.$$

The penalty parameters μ_0, ν_0, μ_N and ν_N are specified in Theorem 3.2.

- v) If $\mathcal{U}_t = 0$, we have a stationary problem and the linear system $LU = \text{RHS}$ must be solved. For the time-dependent cases, we use the method of lines and discretize $U_t + LU = \text{RHS}$ in time using a suitable solver for ordinary differential equations.

Remark 5.1. When we have a hyperbolic problem, step (i) is omitted and step (iv) is modified such that the scheme (11) is used with penalty parameters given in Theorem 2.6.

In the simulations, we are interested in the functional error $\mathbf{E} = J(U) - \mathcal{J}(\mathcal{U})$, where $\mathcal{J}(\mathcal{U}) = \langle \mathcal{G}, \mathcal{U} \rangle$, $J(U) = \langle G, U \rangle_H$ and $G_i(t) = \mathcal{G}(x_i, t)$, but of course also in the solution error \mathbf{e} , where $\mathbf{e}_i(t) = U_i(t) - \mathcal{U}(x_i, t)$. We also investigate the spectra of L , that is the eigenvalues λ_j of L , with $j = 1, 2, \dots, n(N+1)$. Here we are in particular interested in the spectral radius $\rho = \max_j(|\lambda_j|)$ and in $\eta = \min_j(\Re(\lambda_j))$. (For time-dependent problems $\rho \Delta t \lesssim C$ is a crude estimate of the stability regions of explicit Runge-Kutta schemes, and thus ρ can be seen as a measure of stiffness. The eigenvalue with the smallest real part, η , determines how fast a time-dependent solution converges to a steady-state solution, see [14].) Ideally, the penalties are chosen such that ρ is kept small while η is maximized. For steady problems or when using implicit time solvers, other properties (e.g. the condition number) might be of greater interest.

We start by investigating a couple of scalar cases in some detail, then give an example of a system with a solid wall type of boundary condition.

5.1 The scalar case

Consider the scalar advection-diffusion equation,

$$\begin{aligned} \mathcal{U}_t + a\mathcal{U}_x - \varepsilon\mathcal{U}_{xx} &= \mathcal{F}, & x \in [0, 1], \\ \alpha_L \mathcal{U} + \beta_L \mathcal{U}_x &= g_L, & x = 0, \\ \alpha_R \mathcal{U} + \beta_R \mathcal{U}_x &= g_R, & x = 1, \end{aligned} \quad (60)$$

valid for $t \geq 0$, with initial condition $\mathcal{U}(x, 0) = \mathcal{U}_0(x)$ and where $\varepsilon > 0$. Using (30) yields

$$\bar{\mathcal{A}} = \begin{bmatrix} a & -\varepsilon \\ -\varepsilon & 0 \end{bmatrix}, \quad \bar{\mathcal{B}}_L = \begin{bmatrix} \alpha_L & \beta_L \end{bmatrix}, \quad \bar{\mathcal{B}}_R = \begin{bmatrix} \alpha_R & \beta_R \end{bmatrix}.$$

In this case, the factorization of the matrix $\bar{\mathcal{A}}$ can be parameterized as

$$\bar{\mathcal{A}} = \bar{\mathcal{Z}} \bar{\Delta} \bar{\mathcal{Z}}^T = \begin{bmatrix} \frac{a+\omega}{2s_1} & \frac{a-\omega}{2s_2} \\ \frac{-\varepsilon}{s_1} & \frac{-\varepsilon}{s_2} \end{bmatrix} \begin{bmatrix} \frac{s_1^2}{\omega} & 0 \\ 0 & -\frac{s_2^2}{\omega} \end{bmatrix} \begin{bmatrix} \frac{a+\omega}{2s_1} & \frac{a-\omega}{2s_2} \\ \frac{-\varepsilon}{s_1} & \frac{-\varepsilon}{s_2} \end{bmatrix}^T, \quad (61)$$

with $\omega > 0$. In particular, if $\omega = \sqrt{a^2 + 4\varepsilon^2}$ and if $s_{1,2}^2 = \omega(\omega \pm a)/2$, then the above factorization is the eigendecomposition of $\bar{\mathcal{A}}$. The discrete scheme mimicking (60) is

$$\begin{aligned} U_t + aD_1U - \varepsilon D_2U &= F + H^{-1}(\mu_0 e_0 + \nu_0 S^T e_0)(\alpha_L U_0 + \beta_L (SU)_0 - g_L) \\ &\quad + H^{-1}(\mu_N e_N + \nu_N S^T e_N)(\alpha_R U_N + \beta_R (SU)_N - g_R). \end{aligned} \quad (62)$$

To compute the penalty parameters, $\bar{P}_L = \frac{s_1}{\omega}(\alpha_L + \beta_L \frac{a-\omega}{2\varepsilon})$ and $\bar{P}_R = -\frac{s_2}{\omega}(\alpha_R + \beta_R \frac{a+\omega}{2\varepsilon})$ are needed, which we obtain using (59). Theorem 3.2 now yields

$$\begin{aligned} \mu_0 &= \frac{-\frac{a+\omega}{2} - q\varepsilon}{\alpha_L + \beta_L \frac{a-\omega}{2\varepsilon} - q\beta_L}, & \nu_0 &= \frac{-\varepsilon}{\alpha_L + \beta_L \frac{a-\omega}{2\varepsilon} - q\beta_L}, \\ \mu_N &= \frac{\frac{a-\omega}{2} - q\varepsilon}{\alpha_R + \beta_R \frac{a+\omega}{2\varepsilon} + q\beta_R}, & \nu_N &= \frac{\varepsilon}{\alpha_R + \beta_R \frac{a+\omega}{2\varepsilon} + q\beta_R}. \end{aligned} \quad (63)$$

Formally $0 < \omega < \infty$ is necessary (since in the limits $\bar{\mathcal{Z}}$ becomes singular), but as long as the number of imposed boundary condition does not change or the penalty parameters go to infinity, we can allow $0 \leq \omega \leq \infty$. Below we present some special cases:

For Dirichlet boundary conditions we have $\alpha_L = \alpha_R = 1$ and $\beta_L = \beta_R = 0$. In this case the penalty parameters in (63) become

$$\mu_0 = -\frac{a+\omega}{2} - q\varepsilon, \quad \nu_0 = -\varepsilon, \quad \mu_N = \frac{a-\omega}{2} - q\varepsilon, \quad \nu_N = \varepsilon, \quad (64)$$

with $0 \leq \omega < \infty$. Translating the penalty parameters for the advection-diffusion case in [1] to the form used here, it can be seen that they are exactly the same.

With $\alpha_L = \frac{|a|+a}{2}$, $\beta_L = -\varepsilon$ at the left boundary and $\alpha_R = \frac{|a|-a}{2}$, $\beta_R = \varepsilon$ at the right boundary, we have boundary conditions of a low-reflecting far-field type. In this case, the penalty parameters in (63) become

$$\mu_0 = -\frac{\frac{\omega+a}{2} + q\varepsilon}{\frac{\omega+|a|}{2} + q\varepsilon}, \quad \nu_0 = \frac{-\varepsilon}{\frac{\omega+|a|}{2} + q\varepsilon}, \quad \mu_N = -\frac{\frac{\omega-a}{2} + q\varepsilon}{\frac{\omega+|a|}{2} + q\varepsilon}, \quad \nu_N = \frac{\varepsilon}{\frac{\omega+|a|}{2} + q\varepsilon} \quad (65)$$

and we see that in the limit $\omega \rightarrow \infty$, we obtain $\mu_0 = -1$, $\nu_0 = 0$, $\mu_N = -1$ and $\nu_N = 0$. This particular choice corresponds to the penalty $\Sigma = -I$ used in [2, 3] for systems with boundary conditions of far-field type.

Remark 5.2. If $\varepsilon = 0$ in (60) we get the transport equation, and then only one boundary condition should be given instead of two. That means that the derivation of the penalty parameters must be redone accordingly. See [1], where this case is covered.

Remark 5.3. The results can be extended to the case of varying coefficients. Consider the scalar diffusion problem $\mathcal{U}_t - (\varepsilon \mathcal{U}_x)_x = \mathcal{F}$ with Dirichlet boundary conditions, where $\varepsilon(x) > 0$. Following [12], we define a narrow-stencil operator mimicking $\partial/\partial x(\varepsilon \partial/\partial x)$ as

$$D_2^{(\varepsilon)} = H^{-1} \left(-A_S^{(\varepsilon)} + (\varepsilon(1)E_N - \varepsilon(0)E_0)S \right)$$

where $A_S^{(\varepsilon)}$ is symmetric and positive semi-definite. It is assumed that $D_2^{(\varepsilon)} = \varepsilon D_2$ holds when ε is constant. The discrete problem becomes

$$U_t - D_2^{(\varepsilon)} U = F + H^{-1}(\mu_0 e_0 + \nu_0 S^T e_0)(U_0 - g_L) + H^{-1}(\mu_N e_N + \nu_N S^T e_N)(U_N - g_R).$$

The continuous problem is self-adjoint, so for dual consistency $L^* = H^{-1}L^T H = L$ is needed, which is fulfilled if $\nu_0 = -\varepsilon(0)$ and $\nu_N = \varepsilon(1)$. Moreover, using $A_S^{(\varepsilon)} \geq \varepsilon_{\min} A_S$, where $\varepsilon_{\min} = \min_{x \in [0,1]} \varepsilon(x)$, it can be shown that the discretization will be stable if we choose $\mu_0 \leq -\frac{q}{\varepsilon_{\min}} \varepsilon(0)^2$ and $\mu_N \leq -\frac{q}{\varepsilon_{\min}} \varepsilon(1)^2$. (The superconvergence for functionals has been confirmed numerically and the resulting "best" choices of μ_0 and μ_N are similar to what we obtain in the constant case considered below.)

5.1.1 The stationary heat equation with Dirichlet boundary conditions

We consider the heat equation with Dirichlet boundary conditions, i.e. problem (60) with $a = 0$, $\alpha_L, \alpha_R = 1$ and $\beta_L, \beta_R = 0$, which we solve using the scheme (62), with the penalty parameters given by (64), also with $a = 0$. To isolate the errors originating from the spatial discretization, we first look at the steady problem. Thus we let $\mathcal{U}_t = 0$ and solve $-\mathcal{U}_{xx} = \mathcal{F}(x)$ numerically. The resulting quantities ρ and η , the solution error

$\|\mathbf{e}\|_H$ and the functional error $|\mathbf{E}|$ are given (as functions of the parameter ω) in Figure 1. The spectral radius ρ grows with ω , so we do not want $\omega \rightarrow \infty$. On the other hand, the decay rate η shrinks with ω so $\omega \rightarrow 0$ should also be avoided. The errors tend to decrease with increasing ω (the errors naturally varies slightly depending on the choice of \mathcal{F} and \mathcal{G} , but the example in Figure 1 shows a typical behavior). Thus the demand for accuracy is conflicting with the demand of keeping ρ small (the aim to maximize η is met before the aim to minimize the errors and is therefore not a limiting factor in this case). Empirically we have found that a good compromise, which gives small errors without increasing the spectral radius dramatically, is obtained using $\omega \approx q\varepsilon$.

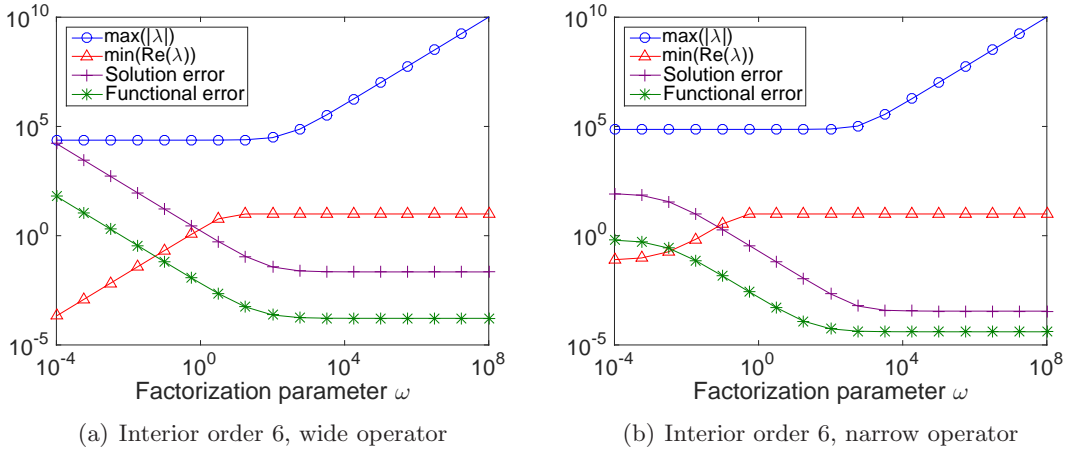


Figure 1: Properties of L and errors when solving $-\mathcal{U}_{xx} = \mathcal{F}(x)$ with Dirichlet boundary conditions. The number of grid points is $N = 64$, the second derivative operator is 6th order accurate in the interior, and is either wide or narrow. Here $\mathcal{U}(x) = \mathcal{G}(x) = \cos(30x)$.

From this example, we make an observation. If we would use the eigenfactorization, we would have $\omega = \sqrt{a^2 + 4\varepsilon^2} = 2$. However, in Figure 1 we see that that choice is not especially good, since the errors then become much larger than if using $\omega = q\varepsilon$, which is approximately 200 and 340, respectively. In some cases, the difference in accuracy is so severe that the choice of factorization parameter ω affects the convergence rate. For the narrow operator with the order (2,0), the errors behave as $\|\mathbf{e}\|_H \sim h^{3/2}$ when using $\omega \sim 1$, whereas we obtain the expected $\|\mathbf{e}\|_H \sim h^2$ when using $\omega \sim 1/h$. Similar behaviors are observed also for narrow operators of higher order, see below.

In Figure 2(a) the errors $\|\mathbf{e}\|_H$ for the operators with interior order 6 are shown. For the narrow scheme, the convergence rate is 4.5 when using $\omega = 2\varepsilon$ and 5.5 when using $\omega = q\varepsilon$. For the wide scheme, the order is 4 in both cases, but the error constant changes. In the 8th order case, Figure 2(b), the convergence rates are not affected, but in the

narrow case the errors are around 2500 times smaller when using $\omega = q\varepsilon$. In this example, the functional errors are not as sensitive to ω as the solution errors. In the 6th order case, the convergence rates are slightly better than the predicted $2p = 6$, both for the wide and the narrow schemes, see Figure 3(a). For the 8th order case, see Figure 3(b), the convergence rates are in all cases higher than $2p = 8$. Thus the derived SAT parameters actually produce superconvergent functionals, also for the narrow operators.

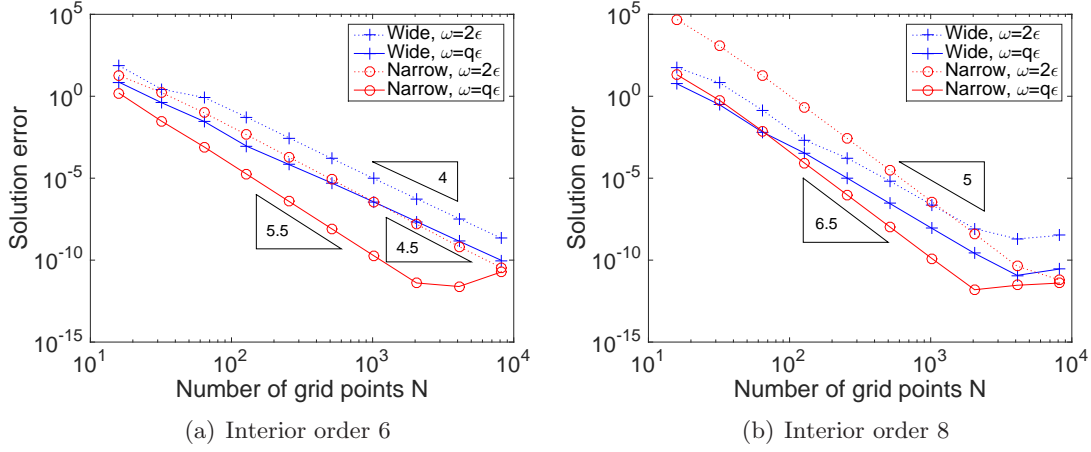


Figure 2: The error $\|\mathbf{e}\|_H$, for $-\mathcal{U}_{xx} = \mathcal{F}(x)$. The exact solution is $\mathcal{U} = \cos(30x)$.

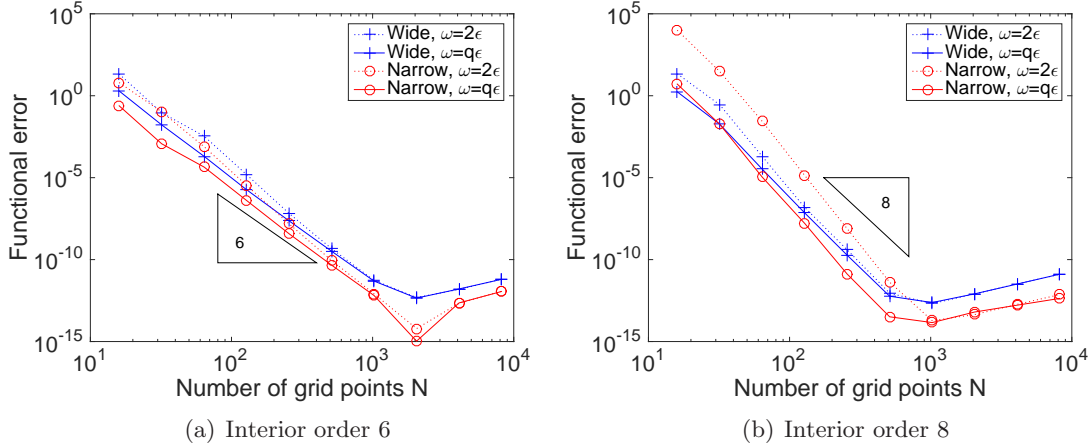


Figure 3: The functional error $|\mathbf{E}|$, using the weight function $\mathcal{G}(x) = \cos(30x)$.

5.1.2 The time-dependent heat equation with Dirichlet boundary conditions

Next, we consider the actual heat equation. We solve $\mathcal{U}_t = \varepsilon \mathcal{U}_{xx} + \mathcal{F}(x, t)$ with $\varepsilon = 0.01$ and the exact solution $\mathcal{U}(x, t) = \cos(30x) + \sin(20x)\cos(10t) + \sin(35t)$. For the time propagation the classical 4th order accurate Runge-Kutta scheme is used, with sufficiently small time steps, $\Delta t = 10^{-4}$, such that the spatial errors dominate. In Figure 4 the errors obtained using the narrow (6,3) order scheme are shown as a function of time.

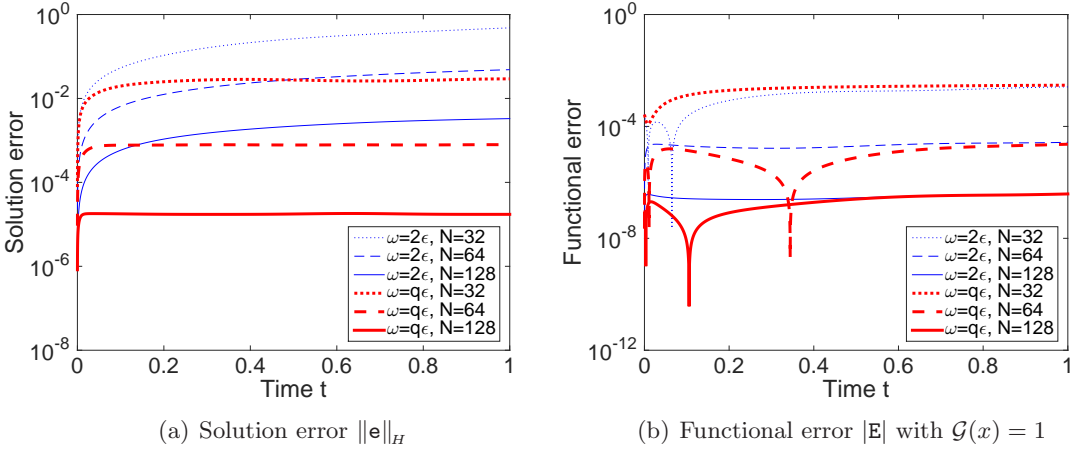


Figure 4: Errors when solving the heat equation using the narrow (6,3) order scheme.

The corresponding spatial order of convergence (at time $t = 1$) is shown in Table 2. The simulations confirm the steady results, namely that both $\omega = 2\varepsilon$ and $\omega = q\varepsilon$ give superconvergent functionals but that choosing the factorization parameter as $\omega \sim \varepsilon/h$ improves the solution significantly compared to when using the eigendecomposition.

N	$\omega = 2\varepsilon$				$\omega = q\varepsilon$			
	$\ e\ _H$	Order	$ E $	Order	$\ e\ _H$	Order	$ E $	Order
32	0.480872	—	0.00258741	—	0.029297	—	0.00297573	—
64	0.048501	3.3096	0.00002704	6.5804	0.000790	5.2121	0.00002315	7.0064
128	0.003307	3.8743	0.00000038	6.1559	0.000017	5.5131	0.00000039	5.9055

Table 2: The errors and convergence rates at $t = 1$ for the narrow (6,3) order scheme.

5.1.3 The heat equation with Neumann boundary conditions

We solve $\mathcal{U}_t = \varepsilon \mathcal{U}_{xx} + \mathcal{F}(x, t)$ again, but this time with Neumann boundary conditions, and the penalty parameters are now given by (63) with $a = 0$, $\varepsilon = 0.01$, $\alpha_L = \alpha_R = 0$ and

$\beta_L = \beta_R = 1$. In contrast to when having Dirichlet boundary conditions, the spectral radius ρ does not depend so strongly on ω and therefore we can let $\omega \rightarrow \infty$. Figure 5 shows the convergence rates for the schemes with interior order 6. The exact solution is $\mathcal{U}(x, t) = \cos(30x)$ and for the time propagation the implicit Euler method, with $\Delta t = 1$, is used (this is more than enough since the chosen \mathcal{U} does not depend on t). We note that the convergence rates behaves similarly to the Dirichlet case. We could also have used $\omega = q\varepsilon$ here, it gives the same convergence rates as $\omega = \infty$.

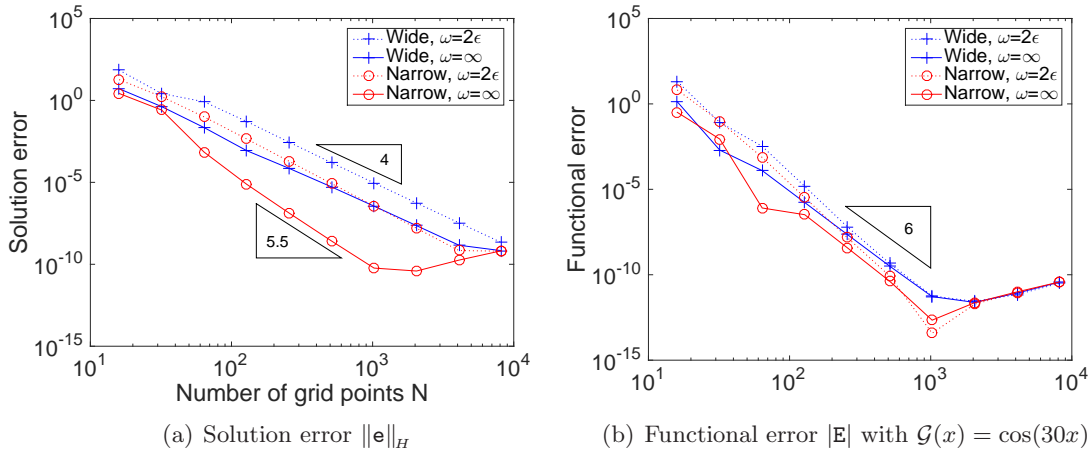


Figure 5: Errors at time $t = 100$ when solving the heat equation with Neumann boundary conditions using the schemes with interior order 6.

5.1.4 The advection-diffusion equation with Dirichlet boundary conditions

For simplicity we consider steady problems again, this time $a\mathcal{U}_x = \varepsilon\mathcal{U}_{xx} + \mathcal{F}$. That is, we solve (60) using the scheme (62), both with omitted time derivatives. The penalty parameters for Dirichlet boundary conditions are given in (64).

First, we take a look at an interesting special case, namely when $\mathcal{F} = 0$. Then the exact solution is $\mathcal{U}(x) = c_1 + c_2 \exp(ax/\varepsilon)$, where the constants c_1 and c_2 are determined by the boundary conditions. For $\varepsilon \ll |a|$ the exact solution forms a thin boundary layer at the outflow boundary, which for insufficient resolution usually leads to oscillations in the numerical solution. This can be handled by upwinding or artificial diffusion (see e.g. [16]). Here we will instead use the free parameter ω in the penalty to minimize the oscillating modes (the so-called π -modes).

We start with the wide second derivatives stencils. The ansatz $U_i = k^i$, inserted into

the interior of the scheme (62), gives (for the second order case) a numerical solution

$$U_i = \tilde{c}_1 + \tilde{c}_2(-1)^i + \tilde{c}_3 k_3^i + \tilde{c}_4 k_4^i, \quad k_{3,4} = \frac{ha}{\varepsilon} \pm \sqrt{\frac{h^2 a^2}{\varepsilon^2} + 1}.$$

Thus there exist two modes with alternating signs, $\tilde{c}_2(-1)^i$ and $\tilde{c}_4 k_4^i$. However, one can show that the choice $\omega = |a|$ leads to $\tilde{c}_2 = 0$ and to \tilde{c}_4 being small enough compared to \tilde{c}_3 such that U_i is monotone. Empirically we have seen that this nice behavior holds also for the wide schemes with higher order of accuracy. In Figure 6 the result using the scheme with interior order 8 is shown. The solution obtained using $\omega = |a|$ shows no oscillations, even though the grid is very coarse. Moreover, this particular choice of factorization gives functional errors almost at machine precision (although it should be noted that this is a special case since $\mathcal{F}(x) = 0$ and $\mathcal{G}(x) = 1$).

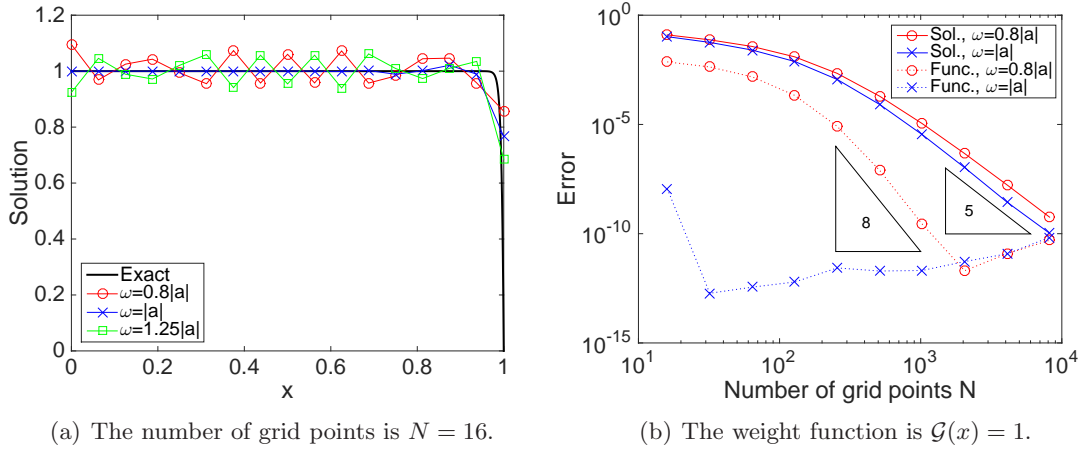


Figure 6: We solve $a\mathcal{U}_x = \varepsilon\mathcal{U}_{xx}$ with $a = 1$, $\varepsilon = 0.005$ using the wide scheme with interior order 8. In a) the solutions, in b) the errors $\|\mathbf{e}\|_H$ and $|\mathbf{E}|$.

For the narrow-stencil schemes, the existence of spurious oscillating modes depends on the resolution. In the second order case, the interior solution is

$$U_i = \tilde{c}_1 + \tilde{c}_2 \left(\frac{1 + ah/(2\varepsilon)}{1 - ah/(2\varepsilon)} \right)^i,$$

which has an oscillating component if $|a|h/(2\varepsilon) > 1$. With very particular choices of the penalty parameter this component can be canceled (for the operators with order (2,0) and (2,1) it is achieved using $\omega = |a|/(1 - \frac{2\varepsilon}{|a|h})$ and $\omega = |a|(1 - \frac{\varepsilon}{|a|h})/(1 - \frac{2\varepsilon}{|a|h})^2$, respectively) such that the numerical solution becomes constant. As soon as $|a|h/(2\varepsilon) < 1$, this mode

should not be canceled anymore, but how to do the transition between the unresolved case and the resolved case is not obvious. For the higher order schemes the ω which cancels the oscillating modes are even more complicated and in some cases negative (i.e. useless). In short, these particular, canceling choices of ω are not worth the effort. Instead, we recommend to use $\omega \approx |a| + q\varepsilon$ for the narrow-stencil operators, see below.

The above results were obtained under the assumption $\mathcal{F} = 0$. Next, we use a forcing function \mathcal{F} such that the exact solution is $\mathcal{U}(x) = \cos(30x)$. The resulting errors, together with ρ and η , are shown in Figure 7 for $a = 1$ and $\varepsilon = 10^{-6}$. Clearly, $\omega \approx |a|$ is still a good choice since the errors are small, ρ is not too large and η is maximal. For $\varepsilon \gg |a|h$ the curves are more similar to those in Figure 1, and $\omega \approx |a| + q\varepsilon$ will be a better choice. In the transition region $\varepsilon \sim |a|h$ we sometimes observe order reduction. This can be seen in Figures 8 and 9 for the schemes with an interior order of accuracy 6. Figure 8 shows the convergence rates when $\varepsilon = 0.1$, which is large enough for the numerical solution to be well resolved. For the narrow scheme, we see an improved convergence rate for the solution error if $\omega = |a| + q\varepsilon$ is used. The functional output converges with $2p = 6$ for all schemes. Figure 9 shows the convergence rates when ε is decreased to 10^{-4} , such that the numerical solution is badly resolved. For all schemes, except the wide scheme with the particular choice $\omega = |a|$, we see a pre-asymptotic order reduction of the functional.

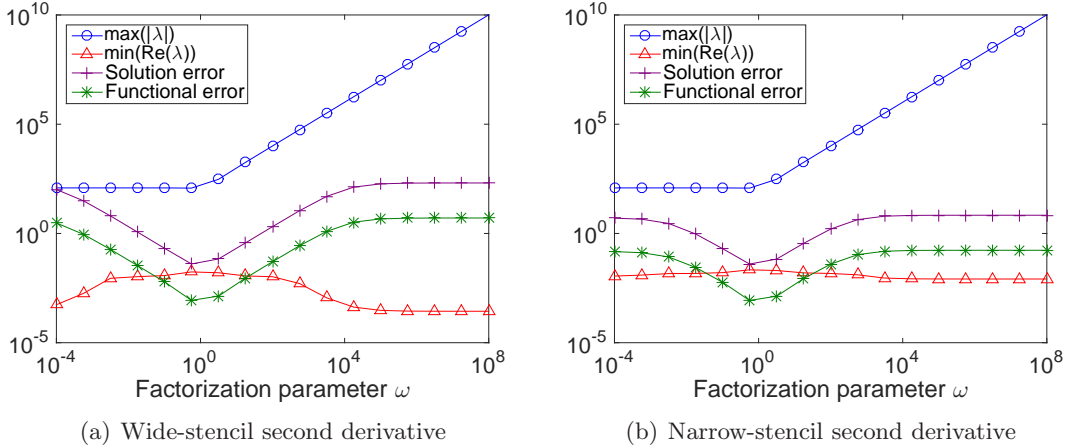


Figure 7: We solve $a\mathcal{U}_x = \varepsilon\mathcal{U}_{xx} + \mathcal{F}(x)$ with Dirichlet boundary conditions and with $\mathcal{U}(x) = \mathcal{G}(x) = \cos(30x)$. The number of grid points is $N = 64$, the interior order is 6.

We conclude that the penalties in Theorem 3.2 yields superconvergent functionals for the advection-diffusion equation with Dirichlet boundary conditions – in the asymptotic limit. In the special case when having the wide scheme with $\omega = |a|$ we even get superconvergent functionals in the troublesome transition region.

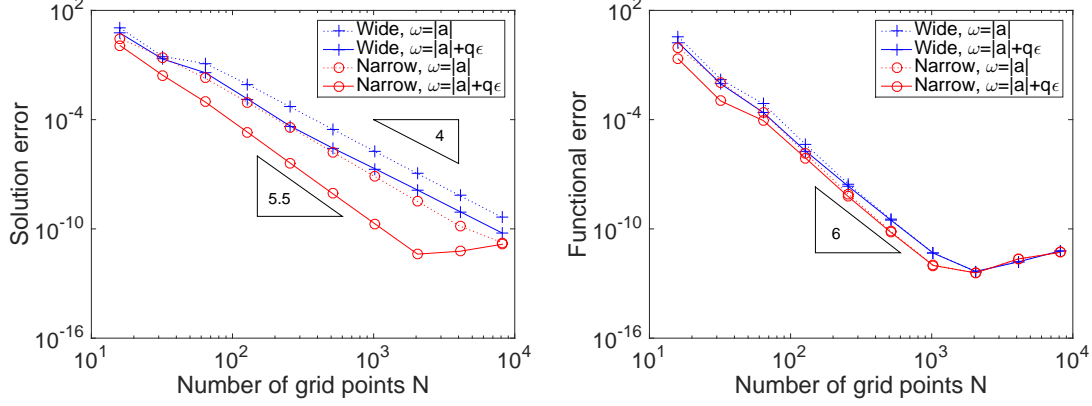


Figure 8: The inner order of accuracy is 6, $\mathcal{U}(x) = \mathcal{G}(x) = \cos(30x)$, $a = 1$ and $\epsilon = 0.1$.

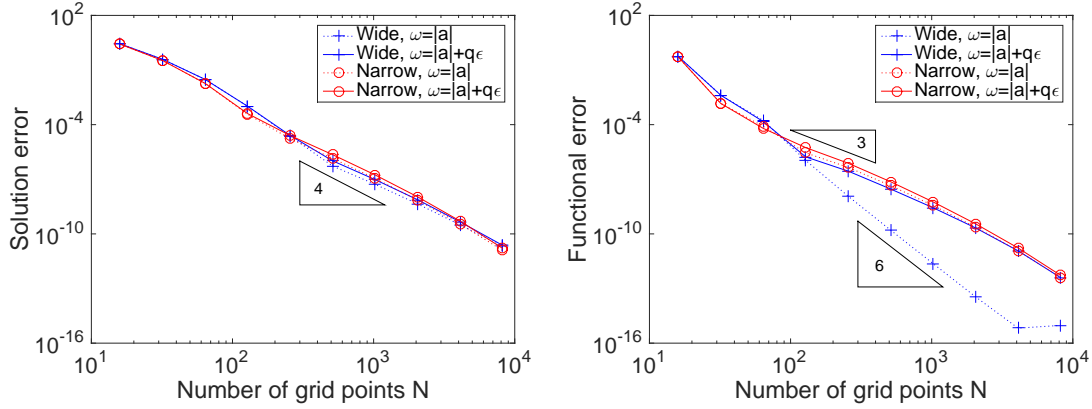


Figure 9: The inner order of accuracy is 6, $\mathcal{U}(x) = \mathcal{G}(x) = \cos(30x)$, $a = 1$ and $\epsilon = 10^{-4}$.

5.1.5 The advection-diffusion equation with far-field boundary conditions

We just comment briefly on the far-field boundary conditions and their corresponding SAT parameters given in (65). If $|a|h/\epsilon$ is large, the quantities ρ , η and the errors barely depend on ω (except if $\mathcal{F} = 0$, then the errors are smaller if $\omega \approx |a|$). For small $|a|h/\epsilon$, large values of ω give smaller errors and slightly larger η , whereas ρ is slightly increased. In this case, the penalty obtained by taking the limit $\omega \rightarrow \infty$, that is $\mu_0 = -1$, $\nu_0 = 0$, $\mu_N = -1$ and $\nu_N = 0$ (corresponding to the penalty used in [2, 3]) is not a bad choice and it has an appealing simplicity. As before, $\omega \approx |a| + q\epsilon$ also gives good results.

5.1.6 Reflections from the scalar case

From what we have seen from the numerical experiments so far, the best choice of the factorization parameter ω is not only dependent on the continuous problem at hand (i.e. the parameters a and ε and the type of boundary conditions), but also on numerical quantities, such as the grid resolution and if the stencils are wide or narrow. In some cases the factorization has almost no impact, sometimes it makes the system at hand extremely ill-conditioned or even changes the order of accuracy of the scheme.

In the scalar case it is rather straightforward to optimize with respect to the single factorization parameter ω , but for systems this task becomes non-trivial and one might have to settle for the factorizations at hand. Nevertheless, we note that the eigendecomposition is not necessarily the best factorization and that it could be worth searching for other options. With that being said, next we consider a system and use nothing but the eigendecomposition for constructing the penalty parameters.

5.2 A fluid dynamics system with solid wall boundary conditions

The symmetrized, compressible Navier–Stokes equations in one dimension ($\Omega = [0, 1]$) with frozen coefficients is given by (28), with

$$\mathcal{A} = \begin{bmatrix} \bar{u} & a & 0 \\ a & \bar{u} & b \\ 0 & b & \bar{u} \end{bmatrix}, \quad \mathcal{E} = \varepsilon \begin{bmatrix} 0 & 0 & 0 \\ 0 & \varphi & 0 \\ 0 & 0 & \psi \end{bmatrix}, \quad \mathcal{U} = \begin{bmatrix} \varrho \\ u \\ T \end{bmatrix},$$

where the constants \bar{u} , a , b , ε , φ and ψ denote suitable physical quantities and where ϱ , u and T are scaled perturbations in density, velocity and temperature. Let $\bar{u} < 0$ and $\varepsilon, \varphi, \psi > 0$. In this case, two boundary conditions should be given at the left boundary and three at the right boundary. We impose solid wall boundary conditions (a perfectly insulated wall) at the left boundary, that is $u(0, t) = T_x(0, t) = 0$. At the right boundary, we impose free stream boundary conditions of Dirichlet type, as $\mathcal{U}(1, t) = \mathcal{U}_\infty$. These boundary conditions give a well-posed problem. The boundary operators are

$$\mathcal{H}_L = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \mathcal{G}_L = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathcal{H}_R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathcal{G}_R = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

These boundary conditions can not be rearranged to the far-field form and therefore the penalty used in [2, 3] can not be applied. We identify $\bar{\mathcal{A}}$, $\bar{\mathcal{B}}_L$ and $\bar{\mathcal{B}}_R$ according to (30), and factorize $\bar{\mathcal{A}}$ using the eigendecomposition. The dual consistent penalty parameters

are now described in (41), with

$$\mathcal{K}_L = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1/(\varepsilon\psi) \end{bmatrix}, \quad \mathcal{K}_R = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

As a comparison, we use the alternative penalty parameters (cf. Method 2 in Remark 4.3)

$$\tilde{\mu}_0 = \begin{bmatrix} -a & 0 \\ 0 & 0 \\ -b & \varepsilon\psi \end{bmatrix}, \quad \tilde{\nu}_0 = \begin{bmatrix} 0 & 0 \\ \varepsilon\varphi & 0 \\ 0 & 0 \end{bmatrix}, \quad \tilde{\mu}_N = \begin{bmatrix} \bar{u} & a & 0 \\ 0 & \bar{u} & 0 \\ 0 & b & \bar{u} \end{bmatrix}, \quad \tilde{\nu}_N = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -\varepsilon\varphi & 0 \\ 0 & 0 & -\varepsilon\psi \end{bmatrix}$$

which give stability (they are chosen such that the boundary terms in (43) are non-positive for zero data) but they do not fulfill the demands for dual consistency.

In the numerical simulations we use the exact solution $\varrho = \cos(7x)$, $u = \sin(13x)$ and $T = \cos(30x)$ and as weight functions we use $\mathcal{G}(x) = [1, 0, 0]^T$, $\mathcal{G}(x) = [0, 1, 0]^T$ and $\mathcal{G}(x) = [0, 0, 1]^T$ (such that one functional output is obtained for each variable). Figure 10 shows the resulting errors when using the schemes with interior order 6. In the wide case, the solutions do not differ much. In the narrow case, the dual consistent solution converges one half order slower than the dual inconsistent one (order 4 for ϱ and 4.5 for u, T compared to 4.5 for ϱ and 5 for u, T), but the result is still as good as in the wide case. Moreover, recall that in the scalar case the order could be improved by choosing another factorization than the eigendecomposition, see Figure 8(a). In Figure 11 we see that the functionals convergence with the expected 6th order for both the dual consistent schemes, whereas the dual inconsistent schemes yield 5th order.

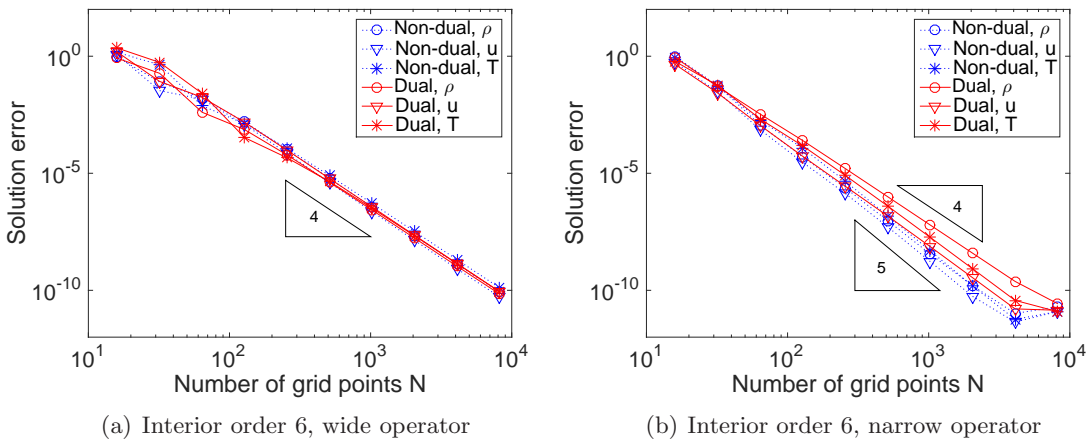


Figure 10: Solution errors, for $\bar{u} = -0.5$, $a = 0.8$, $b = 0.6$, $\varphi = 1$, $\psi = 2$, $\varepsilon = 0.01$.

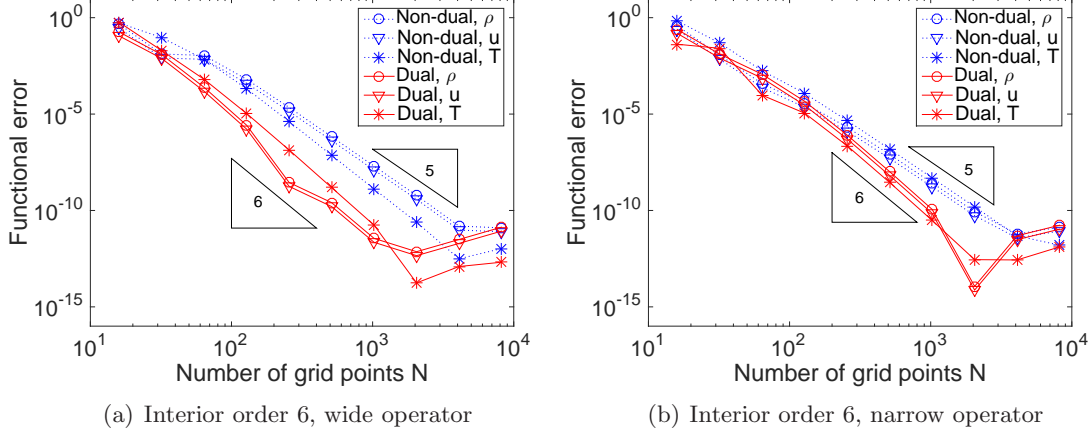


Figure 11: Functional errors, for $\bar{u} = -0.5$, $a = 0.8$, $b = 0.6$, $\varphi = 1$, $\psi = 2$, $\varepsilon = 0.01$.

The diffusion parameter is decreased from $\varepsilon = 0.01$ to $\varepsilon = 10^{-6}$ and the resulting errors are shown in Figures 12 and 13. Now the solution errors obtained using the dual consistent schemes are slightly better than the ones obtained using the dual inconsistent schemes, but the difference is small, see Figure 12.

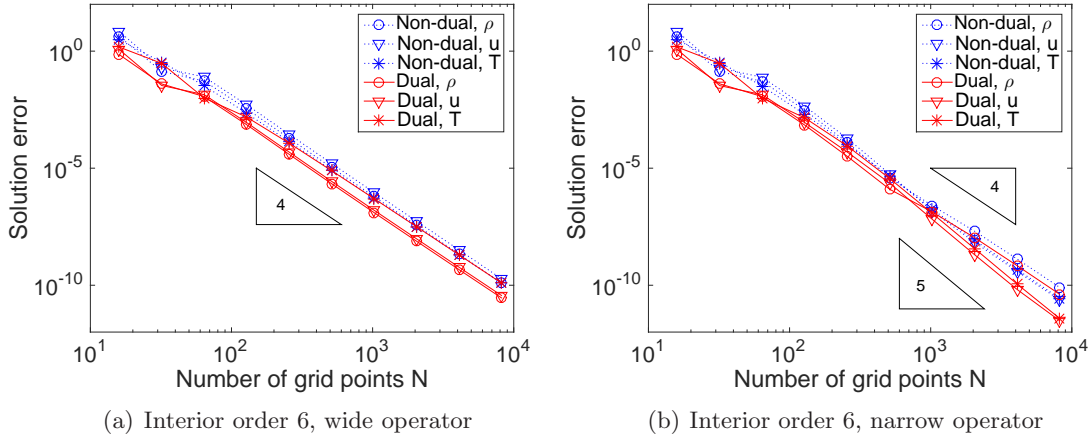


Figure 12: Solution errors, for $\bar{u} = -0.5$, $a = 0.8$, $b = 0.6$, $\varphi = 1$, $\psi = 2$, $\varepsilon = 10^{-6}$.

For the functional errors the difference is more pronounced, see Figure 13. In the wide case, the dual consistent scheme produces a perfect convergence rate of almost 7. This behavior was observed already in the scalar case, when the factorization parameter

was chosen exactly as $\omega = |a|$ (which for small amounts of diffusion is very close to the eigendecomposition). For the narrow-stencil schemes the dual consistent scheme still produces smaller errors than the dual inconsistent scheme, but the order is reduced to 3 (a pre-asymptotic low-order tendency seen already in Figure 9 in the scalar case).

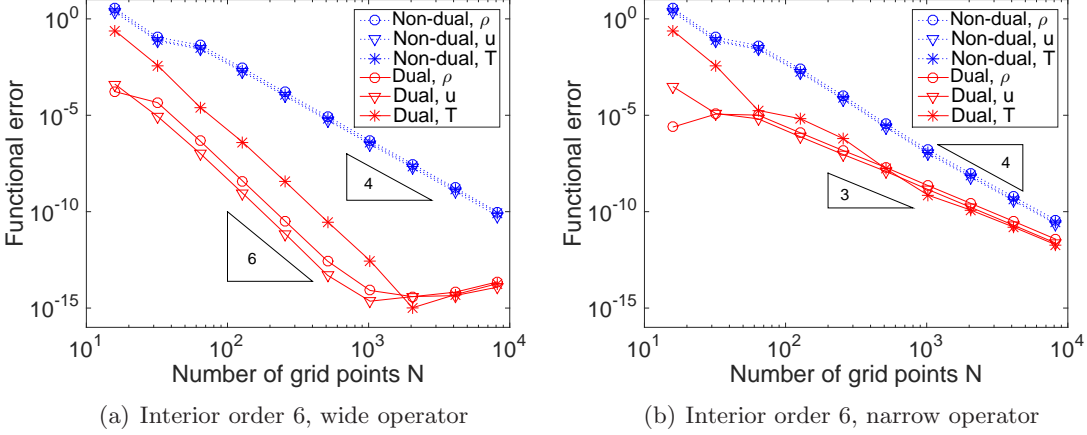


Figure 13: Functional errors, for $\bar{u} = -0.5$, $a = 0.8$, $b = 0.6$, $\varphi = 1$, $\psi = 2$, $\varepsilon = 10^{-6}$.

Extrapolating from the scalar case, we assume that it could be worth searching for better penalty parameters for the narrow-stencil schemes when having diffusion dominated problems. However, for convection dominated problems the wide scheme with a factorization close to the eigendecomposition is hard to beat.

6 Concluding remarks

We use a finite difference method based on summation by parts operators, combined with a penalty method for the boundary conditions (SBP-SAT). Diagonal-norm SBP operators have $2p$ -order accurate interior stencils and p -order accurate boundary closures, which limits the global accuracy of the solution to $p + 1$ (or $p + 2$ for parabolic problems under certain conditions). Recently, it has been shown that SBP-SAT schemes can give functional estimates that are $\mathcal{O}(h^{2p})$. To achieve this superconvergence, the SAT parameters must be carefully chosen to ensure that the discretization is dual-consistent.

We first look at hyperbolic systems and derive stability requirements and duality constraints for the SATs. Then we present a recipe to choose these SAT parameters such that both these (independent) demands are fulfilled. When wide-stencil second derivative operators are used, the results automatically extend to parabolic problems. We generalize

the recipe such that it holds also for narrow-stencil second derivative operators.

The $2p$ order convergence of SBP-SAT functional estimates is confirmed numerically for a variety of scalar examples, as well as for an incompletely parabolic system. For low-diffusion advection-diffusion problems, the superconvergence is sometimes seen first asymptotically. Generally speaking, the narrow-stencil schemes are better for diffusion dominated problems whereas the wide schemes are preferable for advection dominated problems.

In most cases the derived dual consistent SAT parameters have some remaining degree of freedom. The free parameters can be used to improve the accuracy of the primary solution or to tune numerical quantities such as spectral radius, decay rate or condition numbers. Optimal choices within these families are suggested for the scalar problems, however, to do the same for systems is considered a task for the future.

A Reformulation of the first order form discretization

We derive the scheme (31) with penalty parameters (33), using the hyperbolic results.

Step 1: Consider the problem (29), which is a first order system. We represent the solution \bar{U} by a discrete solution vector $\bar{U} = [\bar{U}_0^T, \bar{U}_1^T, \dots, \bar{U}_N^T]^T$, where $\bar{U}_i(t) \approx \bar{U}(x_i, t)$ and discretize (29) exactly as was done in (11) for the hyperbolic case, that is as

$$(I_N \otimes \bar{\mathcal{I}})\bar{U}_t + (I_N \otimes \bar{\mathcal{R}})\bar{U} + (D_1 \otimes \bar{\mathcal{A}})\bar{U} = \bar{F} + (H^{-1}e_0 \otimes \bar{\Sigma}_0)(\bar{\mathcal{B}}_L\bar{U}_0 - g_L) + (H^{-1}e_N \otimes \bar{\Sigma}_N)(\bar{\mathcal{B}}_R\bar{U}_N - g_R). \quad (66)$$

As proposed in Theorem 2.6, we let $\bar{\Sigma}_0 = -\bar{Z}_+\bar{\Delta}_+\bar{P}_L^{-1}$ and $\bar{\Sigma}_N = \bar{Z}_-\bar{\Delta}_-\bar{P}_R^{-1}$.

Step 2: We discretize (28) directly by approximating \mathcal{U} by U and \mathcal{U}_x by \widehat{W} . We obtain

$$U_t + (D_1 \otimes \mathcal{A})U - (D_1 \otimes \mathcal{E})\widehat{W} = F + (H^{-1}e_0 \otimes \sigma_0)(\mathcal{H}_L U_0 + \mathcal{G}_L \widehat{W}_0 - g_L) + (H^{-1}e_N \otimes \sigma_N)(\mathcal{H}_R U_N + \mathcal{G}_R \widehat{W}_N - g_R), \quad (67a)$$

$$(I_N \otimes \mathcal{E})\widehat{W} - (D_1 \otimes \mathcal{E})U = (H^{-1}e_0 \otimes \tau_0)(\mathcal{H}_L U_0 + \mathcal{G}_L \widehat{W}_0 - g_L) + (H^{-1}e_N \otimes \tau_N)(\mathcal{H}_R U_N + \mathcal{G}_R \widehat{W}_N - g_R). \quad (67b)$$

If $\bar{\Sigma}_0 = [\sigma_0^T, \tau_0^T]^T$ and $\bar{\Sigma}_N = [\sigma_N^T, \tau_N^T]^T$, then (67) is a permutation of (66).

Step 3: The scheme in (67) is a system of differential algebraic equations, so we would like to cancel the variable \widehat{W} and get a system of ordinary differential equations instead.

Multiplying (67b) by $\bar{D} = (D_1 \otimes I_n)$ and adding the result to (67a), yields

$$U_t + (D_1 \otimes \mathcal{A})U - (D_1^2 \otimes \mathcal{E})U = F + (H^{-1}e_0 \otimes \sigma_0 + D_1 H^{-1}e_0 \otimes \tau_0)\hat{\chi}_0 \\ + (H^{-1}e_N \otimes \sigma_N + D_1 H^{-1}e_N \otimes \tau_N)\hat{\chi}_N,$$

where

$$\hat{\chi}_0 = \mathcal{H}_L U_0 + \mathcal{G}_L \widehat{W}_0 - g_L, \quad \hat{\chi}_N = \mathcal{H}_R U_N + \mathcal{G}_R \widehat{W}_N - g_R. \quad (68)$$

Next, using the properties in (12), together with the fact that H is diagonal, we compute

$$D_1 H^{-1}e_0 = H^{-1}(-\hat{q}I_N - D_1^T)e_0, \quad D_1 H^{-1}e_N = H^{-1}(\hat{q}I_N - D_1^T)e_N,$$

where \hat{q} is the scalar $\hat{q} = e_0^T H^{-1}e_0 = e_N^T H^{-1}e_N$ given in (35). This yields

$$U_t + (D_1 \otimes \mathcal{A})U - (D_1^2 \otimes \mathcal{E})U = F + \bar{H}^{-1}(e_0 \otimes (\sigma_0 - \hat{q}\tau_0) - D_1^T e_0 \otimes \tau_0)\hat{\chi}_0 \\ + \bar{H}^{-1}(e_N \otimes (\sigma_N + \hat{q}\tau_N) - D_1^T e_N \otimes \tau_N)\hat{\chi}_N, \quad (69)$$

where $\bar{H} = (H \otimes I_n)$. However, the boundary condition deviations $\hat{\chi}_0$ and $\hat{\chi}_N$ still contain \widehat{W} , so we multiply (67b) by $(e_0^T \otimes I_n)$ and $(e_N^T \otimes I_n)$, respectively, to get

$$\mathcal{E}\widehat{W}_0 - \mathcal{E}(\bar{D}U)_0 = \hat{q}\tau_0\hat{\chi}_0, \quad \mathcal{E}\widehat{W}_N - \mathcal{E}(\bar{D}U)_N = \hat{q}\tau_N\hat{\chi}_N. \quad (70)$$

Next, we need boundary condition deviations without \widehat{W} , and define

$$\hat{\xi}_0 = \mathcal{H}_L U_0 + \mathcal{G}_L(\bar{D}U)_0 - g_L, \quad \hat{\xi}_N = \mathcal{H}_R U_N + \mathcal{G}_R(\bar{D}U)_N - g_R.$$

Recall that $\mathcal{G}_{L,R} = \mathcal{K}_{L,R}\mathcal{E}$. Using (70), we can now relate $\hat{\xi}_{0,N}$ above to $\hat{\chi}_{0,N}$ in (68) as

$$\hat{\xi}_0 = (I_{m_+} - \hat{q}\mathcal{K}_L\tau_0)\hat{\chi}_0, \quad \hat{\xi}_N = (I_{m_-} - \hat{q}\mathcal{K}_R\tau_N)\hat{\chi}_N, \quad (71)$$

where I_{m_+} and I_{m_-} are identity matrices of sizes corresponding to the number of positive (m_+) and negative (m_-) eigenvalues of $\bar{\mathcal{A}}$, respectively. Inserting $\hat{\chi}_{0,N}$ from (71) into (69) allows us to finally write the scheme without any \widehat{W} terms and we obtain (31), with

$$\hat{\mu}_0 = (\sigma_0 - \hat{q}\tau_0)(I_{m_+} - \hat{q}\mathcal{K}_L\tau_0)^{-1}, \quad \hat{\nu}_0 = -\tau_0(I_{m_+} - \hat{q}\mathcal{K}_L\tau_0)^{-1}, \\ \hat{\mu}_N = (\sigma_N + \hat{q}\tau_N)(I_{m_-} - \hat{q}\mathcal{K}_R\tau_N)^{-1}, \quad \hat{\nu}_N = -\tau_N(I_{m_-} - \hat{q}\mathcal{K}_R\tau_N)^{-1}. \quad (72)$$

From Step 1 and 2 we know that

$$\begin{bmatrix} \sigma_0 \\ \tau_0 \end{bmatrix} = - \begin{bmatrix} \bar{Z}_1 \bar{\Delta}_+ \bar{P}_L^{-1} \\ \bar{Z}_2 \bar{\Delta}_+ \bar{P}_L^{-1} \end{bmatrix}, \quad \begin{bmatrix} \sigma_N \\ \tau_N \end{bmatrix} = \begin{bmatrix} \bar{Z}_3 \bar{\Delta}_- \bar{P}_R^{-1} \\ \bar{Z}_4 \bar{\Delta}_- \bar{P}_R^{-1} \end{bmatrix},$$

where $\bar{Z}_{1,2,3,4}$ are given in (34). Inserting the above relation into (72), we obtain the penalty parameters presented in (33).

B Motivation of Proposition 4.1

In Proposition 4.1 we claim that the inverse of $\tilde{A}_S = A_S + \delta E_j$ is $J/\delta + K_j$. We motivate this below, for $j = 0$. First, we name the parts of A_S and present the structure of K_0 as

$$A_S = \begin{bmatrix} a & \vec{a}^T \\ \vec{a} & \bar{A} \end{bmatrix}, \quad K_0 = \begin{bmatrix} 0 & \vec{0}^T \\ \vec{0} & \bar{A}^{-1} \end{bmatrix}.$$

Since A_S consists of consistent difference operators, it does not "see" constants. Therefore, $A_S J = 0$ (since J is an all-ones matrix) and $\vec{a} + \bar{A} \vec{1} = \vec{0}$, where $\vec{1} = [1, 1, \dots, 1]^T$. Moreover, due to the special structure of K_0 , we know that $E_0 K_0 = 0$. Thus we have

$$(A_S + \delta E_0)(J/\delta + K_0) = A_S K_0 + E_0 J = \begin{bmatrix} 1 & \vec{a}^T \bar{A}^{-1} + \vec{1}^T \\ \vec{0} & \bar{I} \end{bmatrix} = I.$$

The simplest possible example is the narrow (2,0) order operator in Table 1, specified by

$$D_2 = \frac{1}{h^2} \begin{bmatrix} 0 & 0 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 0 & 0 \end{bmatrix}, \quad \text{with} \quad H = h \begin{bmatrix} 1/2 & & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & & 1/2 \end{bmatrix}. \quad (73)$$

Using (37) and the above structure of K_0 , respectively, we obtain

$$S = \frac{1}{h} \begin{bmatrix} -1 & 1 & & & \\ \times & \times & \times & \times & \times \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \times & \times & \times & \times & \times \\ & & & -1 & 1 \end{bmatrix}, \quad K_0 = h \begin{bmatrix} 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 1 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & \dots & N-1 & N-1 \\ 0 & 1 & \dots & N-1 & N \end{bmatrix}.$$

The interior rows of S are marked by \times 's because they are unknown. Next, we compute

$$\tilde{M}^{-1} = S \tilde{A}_S^{-1} S^T = S (J/\delta + K_0) S^T = \frac{1}{h} \begin{bmatrix} 1 & \times & \dots & \times & 0 \\ \times & \times & \dots & \times & \times \\ \vdots & \vdots & & \vdots & \vdots \\ \times & \times & \dots & \times & \times \\ 0 & \times & \dots & \times & 1 \end{bmatrix}.$$

Just as A_S , the difference stencils in the first and last row of S do not "see" J . Therefore, the corner elements of \tilde{M}^{-1} only depend on K_0 and are independent of δ . We conclude

that when M is singular and S is non-singular the constants in (40) can be computed using (58). In this case we get $q_0 = q_N = 1/h$ and $q_c = 0$, such that $q = 1/h$.

In addition to the operator discussed above, we use the diagonal-norm operators in [13]. For the higher order accurate operators found in [13], q varies with N . For example, for the narrow (4,2) order accurate operator, we have

N	$q_0 h$	$q_c h$	$q h$
8	3.986350339808304	0.000041141179445	3.986391480987749
9	3.986350339313381	0.000002953803786	3.986353293117168
10	3.986350339310830	0.000000212073570	3.986350551384400
11	3.986350339310817	0.000000015226197	3.986350354537014
12	3.986350339310817	0.000000001093192	3.986350340404008

Since the values do not differ so much, it is practical to use the largest value, the one for $N = 8$, regardless of the number of grid points.

References

- [1] J. Berg and J. Nordström. Superconvergent functional output for time-dependent problems using finite differences on summation-by-parts form. *Journal of Computational Physics*, 231(20):6846–6860, 2012.
- [2] J. Berg and J. Nordström. On the impact of boundary conditions on dual consistent finite difference discretizations. *Journal of Computational Physics*, 236:41–55, 2013.
- [3] J. Berg and J. Nordström. Duality based boundary conditions and dual consistent finite difference discretizations of the Navier–Stokes and Euler equations. *Journal of Computational Physics*, 259:135–153, 2014.
- [4] M. H Carpenter, J. Nordström, and D. Gottlieb. A stable and conservative interface treatment of arbitrary spatial accuracy. *Journal of Computational Physics*, 148(2):341–365, 1999.
- [5] S. Eriksson and J. Nordström. Analysis of the order of accuracy for node-centered finite volume schemes. *Applied Numerical Mathematics*, 59(10):2659–2676, 2009.
- [6] D. C. Del Rey Fernández, J. E. Hicken, and D. W. Zingg. Review of summation-by-parts operators with simultaneous approximation terms for the numerical solution of partial differential equations. *Computers & Fluids*, 95:171 – 196, 2014.
- [7] B. Gustafsson, H.-O. Kreiss, and J. Oliger. *Time-Dependent Problems and Difference Methods*. John Wiley & Sons, Inc., 2013.

- [8] J. E. Hicken. Output error estimation for summation-by-parts finite-difference schemes. *Journal of Computational Physics*, 231(9):3828–3848, 2012.
- [9] J. E. Hicken and D. W. Zingg. Superconvergent functional estimates from summation-by-parts finite-difference discretizations. *SIAM Journal on Scientific Computing*, 33(2):893–922, 2011.
- [10] J. E. Hicken and D. W. Zingg. Summation-by-parts operators and high-order quadrature. *Journal of Computational and Applied Mathematics*, 237(1):111–125, 2013.
- [11] H.-O. Kreiss and J. Lorenz. *Initial-boundary value problems and the Navier-Stokes equations*. Academic Press, New York, 1989.
- [12] K. Mattsson. Summation by parts operators for finite difference approximations of second-derivatives with variable coefficients. *Journal of Scientific Computing*, 51(3):650–682, 2012.
- [13] K. Mattsson and J. Nordström. Summation by parts operators for finite difference approximations of second derivatives. *Journal of Computational Physics*, 199(2):503–540, 2004.
- [14] J. Nordström, S. Eriksson, and P. Eliasson. Weak and strong wall boundary procedures and convergence to steady-state of the Navier-Stokes equations. *Journal of Computational Physics*, 231(14):4867–4884, 2012.
- [15] J. Nordström and M. Svärd. Well-posed boundary conditions for the Navier-Stokes equations. *SIAM Journal on Numerical Analysis*, 43(3):1231–1255, 2005.
- [16] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical Mathematics*. Springer-Verlag, 2000.
- [17] B. Strand. Summation by parts for finite difference approximation for d/dx . *Journal of Computational Physics*, 110(1):47 – 67, 1994.
- [18] M. Svärd and J. Nordström. On the order of accuracy for difference approximations of initial-boundary value problems. *Journal of Computational Physics*, 218(1):333–352, 2006.
- [19] M. Svärd and J. Nordström. Review of summation-by-parts schemes for initial-boundary-value problems. *Journal of Computational Physics*, 268:17–38, 2014.